生命科学名著

C.M. 弗雷泽
〔美〕T. D. 里 德 主编
K. E. 纳尔逊
许朝晖
喻子牛 等 译

微生物基因组

MICROBIAL GENOMES



科学出版社
www.sciencep.com

(Q-1600.0101)

基因组工程开创了微生物学研究的新领域，它可以探索数以百计的基因组和发现几百万个新基因。在本书中，基因组的一批开拓者——很多来自基因组研究所(TIGR)——对基因组的现状进行了评价。通过28个发人深思的章节，作者们描述了一些最通用的计算方法，将它们用于病原微生物研究；阐述了基因组怎样用于重建微生物界的发展历史和动力论；讨论了微生物代谢途径、细胞周期、微生物进化、后基因组以及疫苗的开发。此外还探讨了微阵列技术、表达分析以及基因组在药物开发中的作用。

《微生物基因组》全面而有侧重，不仅把现代微生物基因组的许多主要成果汇集在一起，而且综合评述了测序工作在帮助我们理解基因组结构、进化和生物学方面所起的作用。

特　点

- 根据基因组测序论述了微生物的代谢和进化
- 应用微阵列、蛋白质组和基因组学发现疫苗和药物
- 一些章节探讨了微生物注释和感染疾病生物信息学
- 选登了微生物基因组的主要贡献者的部分论文
- 绘制了多个全基因组代谢途径的彩图
- 从历史角度探索性地阐述了基因组的发展进程

主译简介



许朝晖　博士，2000年在华中农业大学获微生物学理学博士学位，1997～1999年赴韩国科学技术院进行合作研究，2000～2002年在美国加州大学河滨分校进行博士后研究。从事新月柄杆菌鞭毛形成及细胞发育机制的研究。发表学术论文多篇，申请专利2项。



喻子牛　教授，博士生导师，国家级专家。武汉华中农业大学农业微生物学国家重点实验室学术委员会主任，微生物农药国家工程研究中心主任。1964～1967年就读陈华癸院士的硕士研究生，1968年以来一直从事苏云金芽孢杆菌的基础研究和应用开发。承担国内外课题40多项。获国家和省部级奖励10项，申请发明专利8项。发表学术论文200多篇，主编专著、教材等8部。培养硕士和博士研究生及博士后108人。



销售分类建议：生物/分子生物学，微生物学

生命科学编辑部
联系电话：010-64012501
http://www.lifescience.com.cn
e-mail:spbio@163.net

ISBN 7-03-015664-1
定　价：75.00 元

微生物基因组

〔美〕C.M. 弗雷泽　T.D. 里德　K.E. 纳尔逊　主编

许朝晖　喻子牛　等　译

科 学 出 版 社

北 京

图字:01-2004-6737

内 容 简 介

本书是由微生物基因组学开创者们撰写,重点介绍了10年来微生物学转向全基因组序列研究的进展,包括微生物基因组学的历史、作为基因组学工具的生物信息学、核心功能、微生物基因组的进化、微生物基因组的调查和基因组数据库的应用共6个部分。所有内容均涉及本学科前沿,作者们现身说法,深入浅出,既能使初涉微生物基因组学领域的研究生们感兴趣,又能使在微生物学和基因组学方面有造诣的专家们参考,是微生物学专著中的精品。本书可供从事微生物学、基因组学、病理学、生态学、酶学、蛋白质组学、植物病理学等领域的研究生、教师和研究人员阅读。

Microbial Genomes
by Claire M. Fraser, Timothy D. Read & Karen E. Nelson
The original English language work has been published by HUMANA PRESS
Totowa, New Jersey, U.S.A.

图书在版编目(CIP)数据

微生物基因组/〔美〕弗雷泽等主编;许朝晖等译.—北京:科学出版社,2006

ISBN 7-03-015664-1

Ⅰ.微… Ⅱ.①弗…②许… Ⅲ.微生物-基因组-研究 Ⅳ.Q933

中国版本图书馆CIP数据核字(2005)第060598号

责任编辑:李 锋 盖 宇 王 静 李军德/责任校对:钟 洋
责任印制:钱玉芬/封面设计:王 浩

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2006年1月第 一 版 开本:787×1092 1/16
2006年1月第一次印刷 印张:30 1/2 插页:4
印数:1—3 000 字数:694 000

定价:75.00元

(如有印装质量问题,我社负责调换〈双青〉)

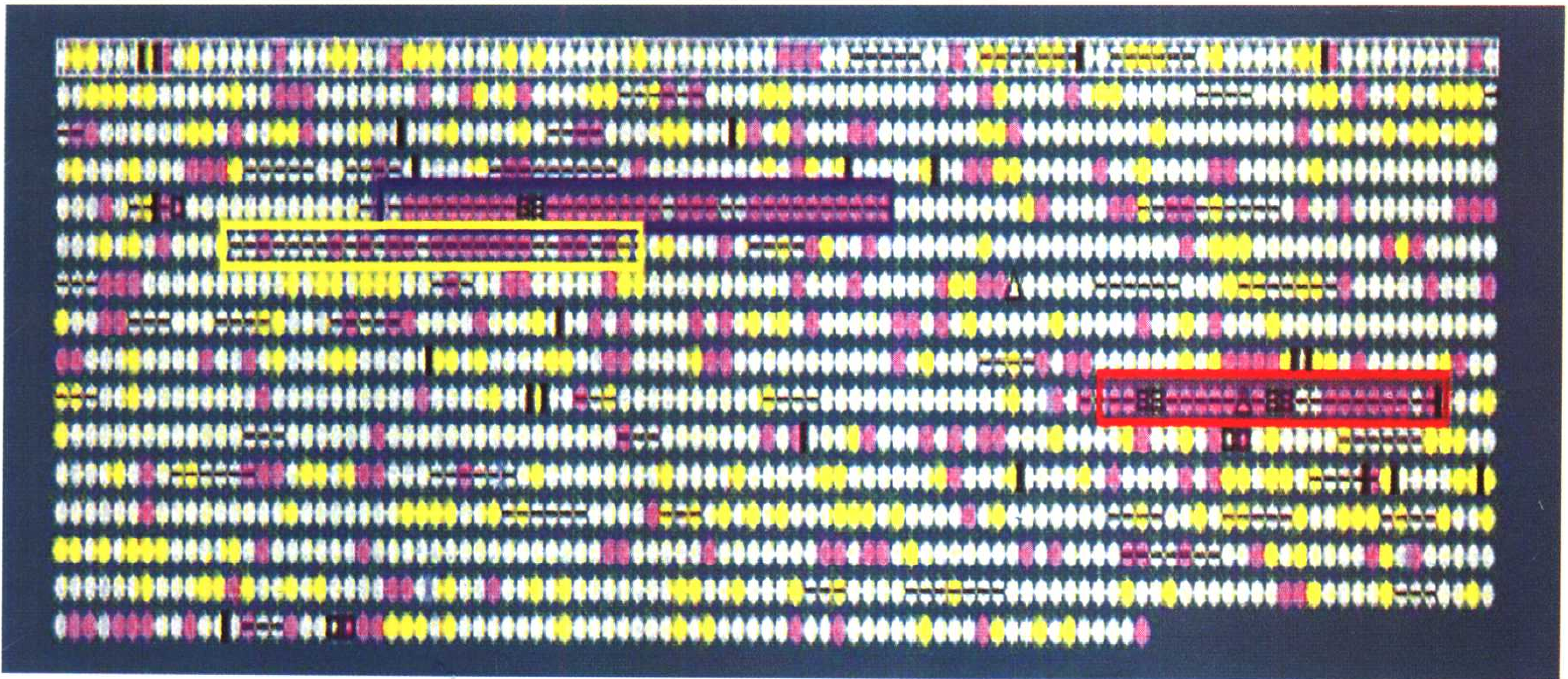


图 4-1 IslandPath 对幽门螺杆菌菌株 26695 全基因组进行分析的输出结果，列出了所有已知致病岛和可能的致病岛。每个圈代表一个预测编码蛋白的可读框。圈的颜色表明 G + C 百分比 (黄色代表高于选定的高阈值，粉红色代表低于选定的低阈值，绿色为两者之间)。圈中的横线代表二核苷酸偏向性 [28]， 竖杆表示转运核糖核酸和核糖体核糖核酸 (黑色为转运核糖核酸，紫色为核糖体核糖核酸，深蓝色为转运核糖核酸和核糖体核糖核酸)。黑色方块为已知的或可能的转座酶基因。黑色三角表示已知的或可能的整合酶基因。具有几项这样特征的区域可能为基因组岛。该菌株中 3 个已知的或可能的岛以彩色框表示：黄框，CAG 致病岛；蓝框，含有 virB 基因的同源序列但在幽门螺杆菌菌株 J99 中不存在的区域；红框内的基因在菌株 J99 和 26995 中不尽相同。请注意，有二核苷酸偏向性的大片段区域 (长横线) 与已知的或可能的基因组岛有很好的相关性。有关 IslandPath 的使用在网上以沙门氏菌 (*Salmonella*) 为例另有介绍 (http: // www.pathogenomics. sfu. ca/islandpath/)。由 IslandPath 新近找出已知的或可能的岛均被标示。

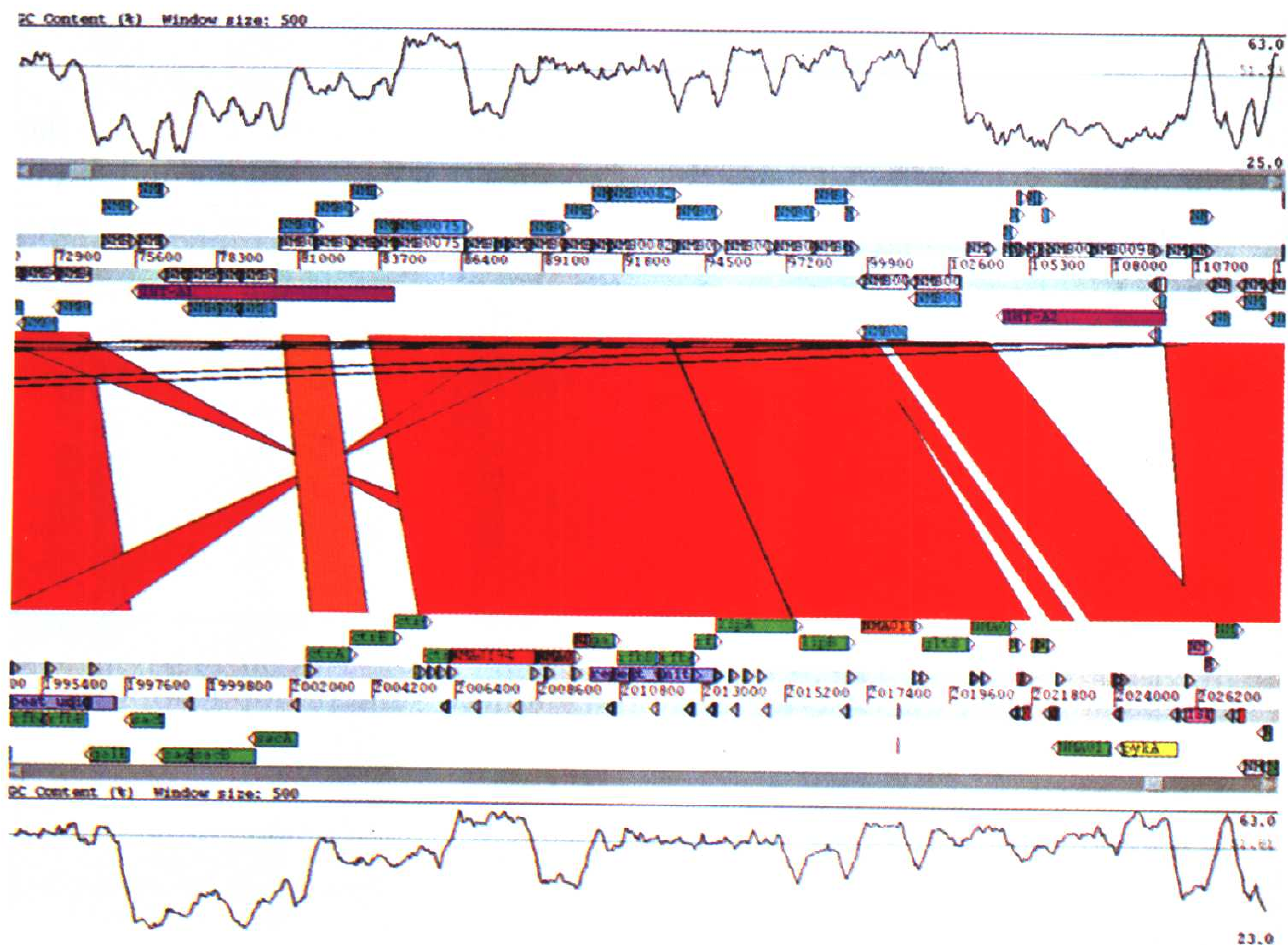


图 4-2 ACT 分析结果窗口。ACT 可以比较不同病原基因组或病原基因组和非病原基因组。每个红色 / 粉红色区域相当一个 BLAST 搜索结果，红色代表较好吻合，白色 / 浅粉红色代表低值吻合。红色 / 粉红色区域上下方序列分别为参考序列 (上方) 的正向和反向序列及查询序列 (下方) 的正向和反向序列；每个方向都有三个可读框。G + C 百分比用斜窗口表示。该图为脑膜炎奈瑟氏球菌菌株 MC58 (血清型 B) 和菌株 Z2491 (血清型 A)。脑膜炎奈瑟氏球菌 A 和 B 血清型有不同的毒力表型。采用这种工具，可将病原菌基因组中与特定疾病表型有关的区域找出来，有利于发现与微生物致病作用有关的序列。

图 6-1 根据恶臭假单胞菌 (*Pseudomonas putida*) 基因组信息重建的代谢途径。复制经《环境微生物学》同意。

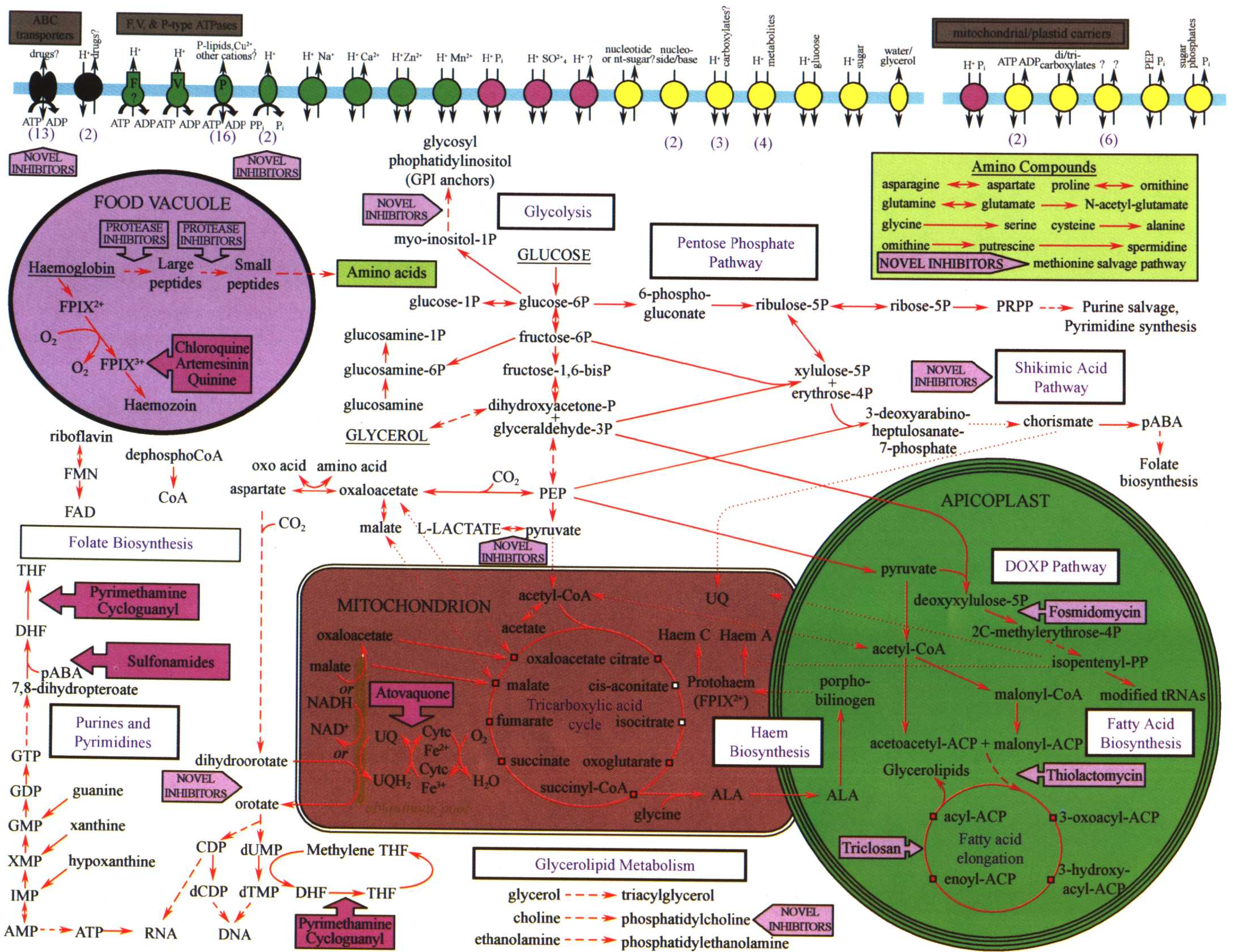


图 6-3 依照恶性疟原虫基因组而重建的代谢途径图。经《自然》杂志允许而重印。

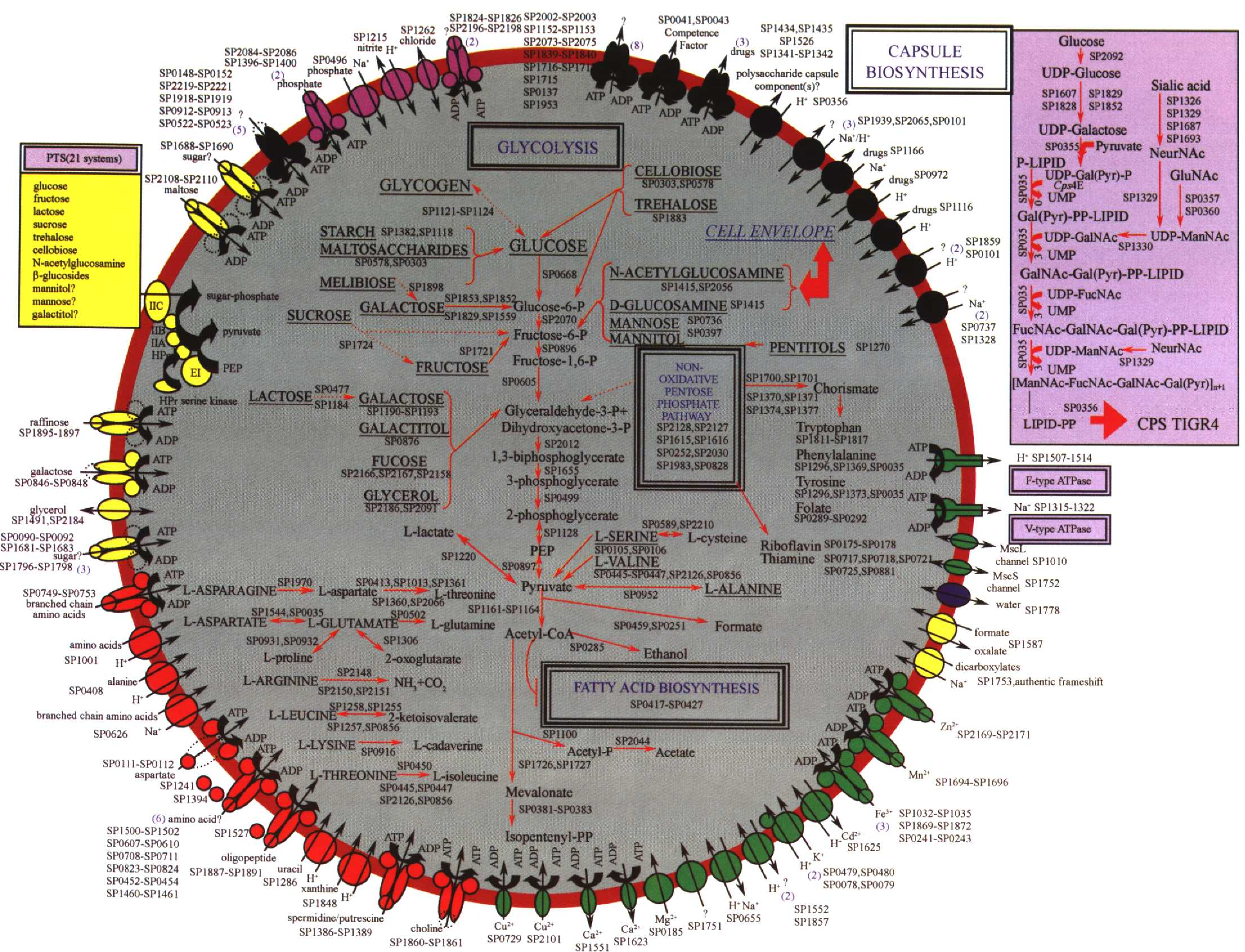


图 7 - 3 肺炎链球菌 (*Streptococcus pneumoniae*) 转运和代谢模型。图中所示为产能、有机化合物的代谢及其荚膜的合成。转运蛋白按照如下所示的底物特异性进行分类：无机阳离子 (绿色)、无机阴离子 (粉色)、碳水化合物/羧酸盐 (黄色)、氨基酸/多肽/胺/嘌呤/嘧啶 (红色)、药物排出和其他 (黑色)。问号表示所转运的底物不能确定。溶质的输出和输入用穿过转运蛋白的箭头线表示。转运蛋白的能量耦合机制按如下所示：利用蛋白通道溶质的转运用双向箭头线表示；次级转运蛋白用两个箭头线表示，分别代表溶质和伴随转运的离子；ATP 驱动转运蛋白用ATP水解反应标明；未知能量耦合机制的转运蛋白用单箭头线标明。组分未知的功能类似多亚基复合体的转运蛋白系统用虚线表示。当具有类似预测底物的多个同源转运蛋白 (multiple homologues transporter) 存在时，该类型转运蛋白的数目在圆括号内标明。系统的基因编号 (SPxxxx) 在每个代谢途径或转运蛋白旁边注明；那些用破折号间隔开的代表一系列连续基因。

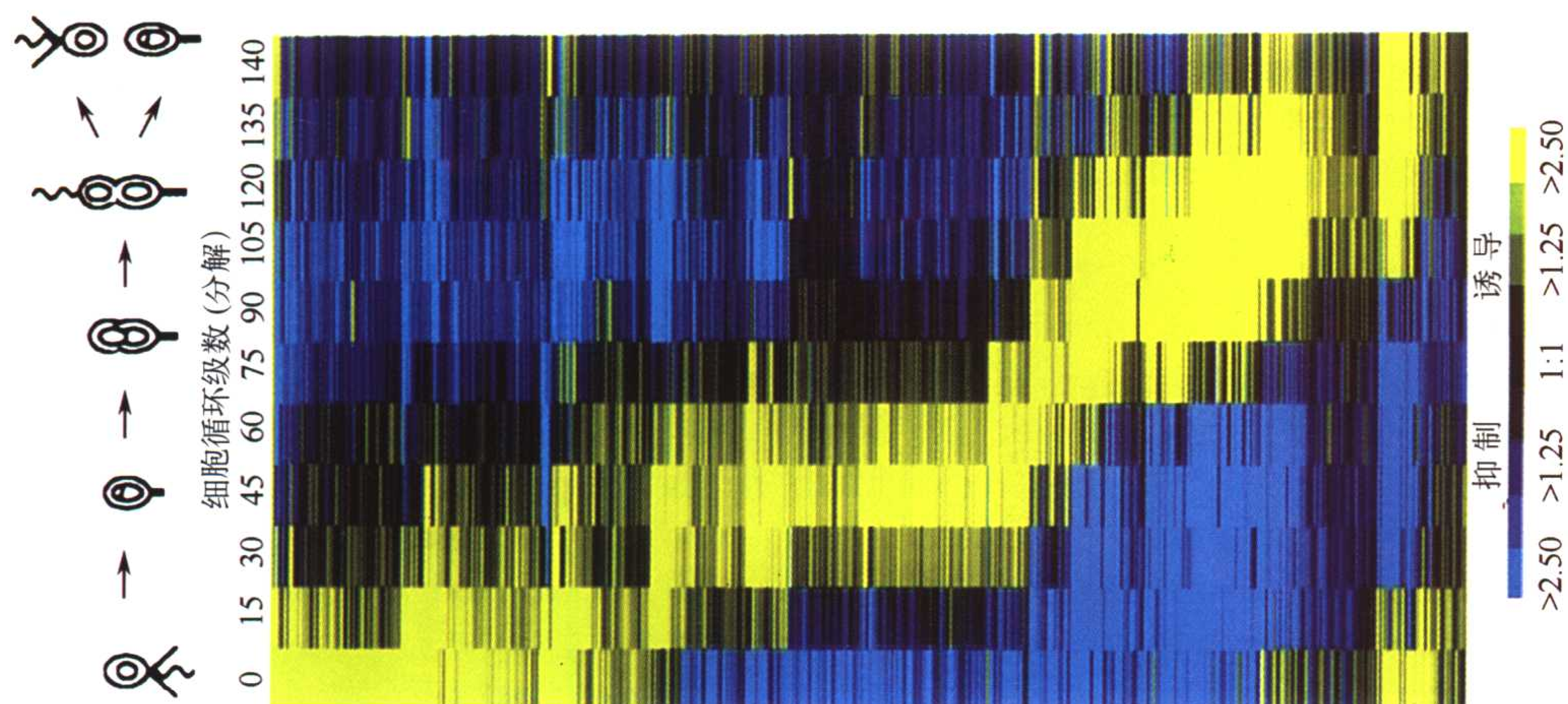


图 8-2 受细胞周期调控的基因表达。发现的 553 个受细胞周期调控的柄杆菌基因的表达谱以颜色表示。图下方的颜色设计代表 RNA 的相对水平。每个基因的表达谱都是从左到右进行。图上方是细胞周期的示意图以及以分钟数表示的细胞周期进展的时间。（根据许可，依照参考文献[9] 重印。American Association for the Advancement of Science 2000 版权所有）

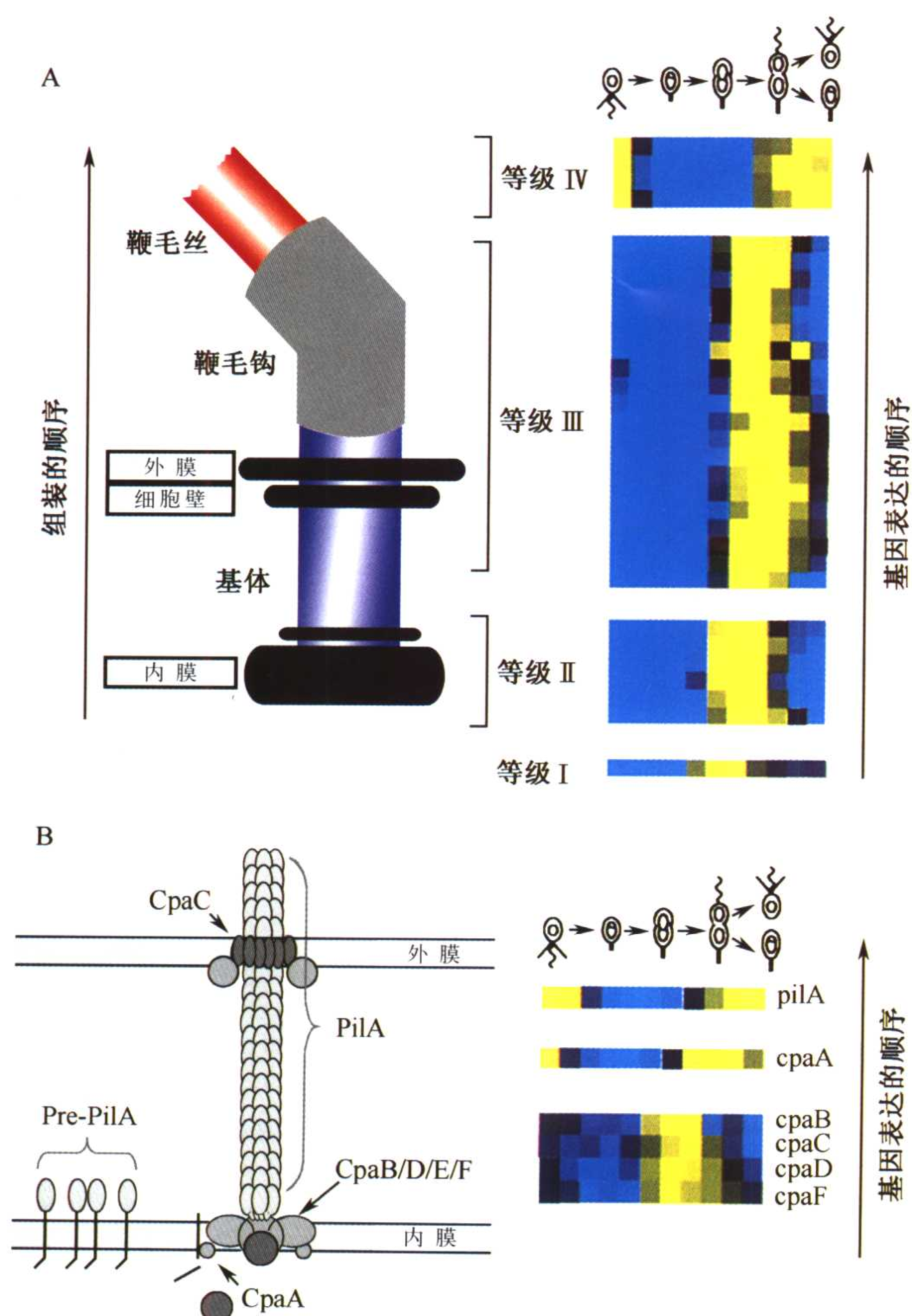


图 8-3 用依次转录的方法控制多蛋白结构的组装。A. 柄杆菌鞭毛组装所必需基因的表达谱被标示在鞭毛示意图旁边。左右两旁箭头分别代表鞭毛组装时间和鞭毛基因表达时间，二者是同步的（见正文）。B. 纤毛生成基因的表达谱被标示在纤毛示意图旁边。这些纤毛基因的表达顺序暗示，同鞭毛基因一样，纤毛基因的表达时间可能对纤毛的正确组装起着至关重要的作用。

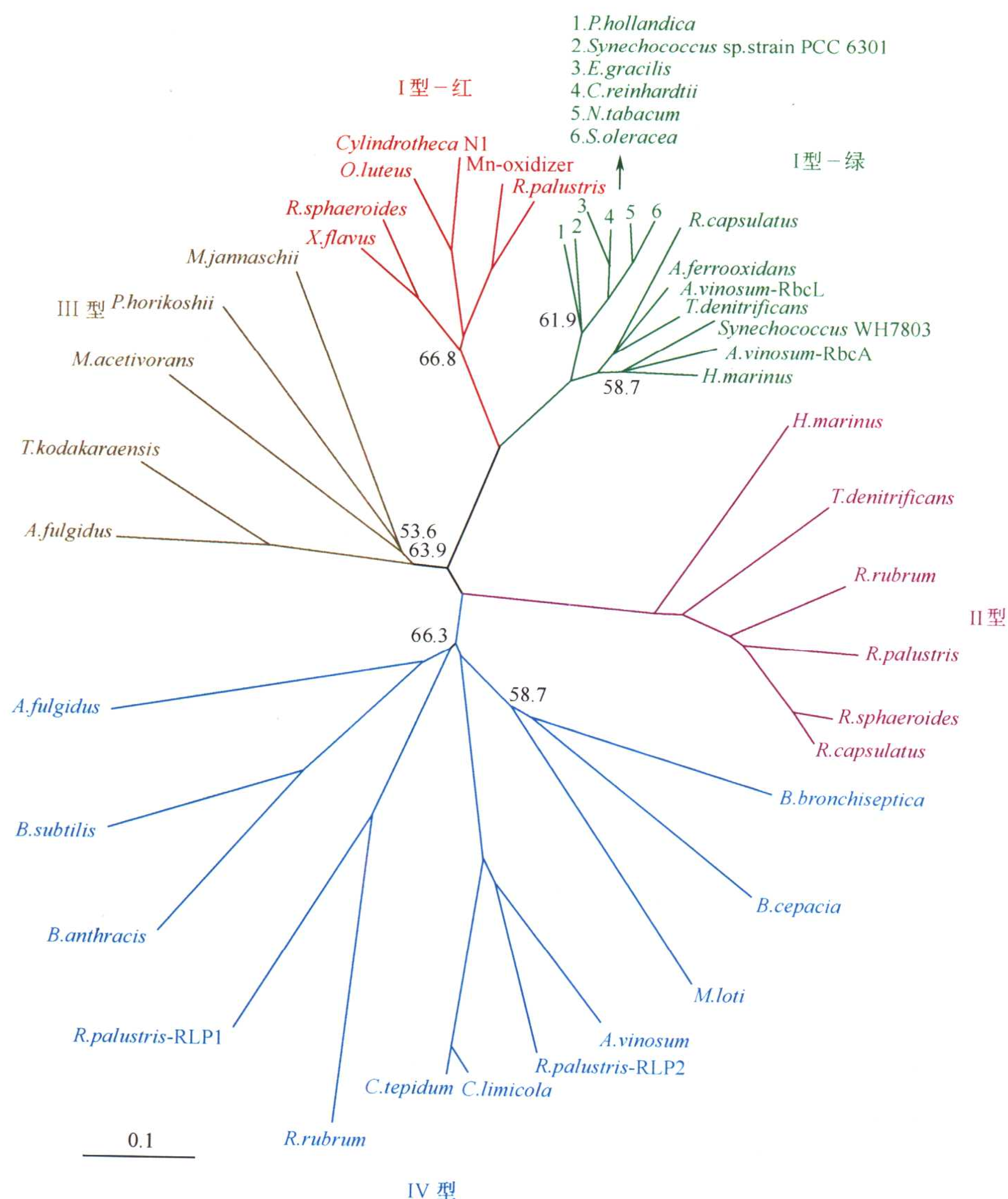


图 14-7 无根相邻种系发生树显示了真正 RubisCo 系列（Ⅰ型，红和绿；Ⅱ型，紫色；Ⅲ型，青铜色）与 RubisCO 类似蛋白系列（Ⅳ型，蓝绿色）间的相互关系。节点位置的引导指令值显示，在 1000 次试验中特定节点出现的次数的百分比。仅标出小于70%引导指令的节点。刻度条代表每个位点替换值为0.1。在文中其他地方未出现过的微生物全名包括：*A. ferroxidans*, *Acidithiobacillus ferroxidans*；*B. anthracis*，炭疽芽胞杆菌 (*Baccilus anthracis*)；*B. cepacia*，洋葱伯克霍尔德氏菌 (*Burkholderia cepacia*)；*B. sbutilis*，枯草芽孢杆菌 (*Bacillus subtilis*)；*B. bronchiseptica*，支气管炎博德特氏菌 (*Bordetella bronchiseptica*)；*C. reinhardtii*，莱茵衣藻 (*Chlamydomonas reinhardtii*)；*E.* gracilis，纤细裸藻 (*Euglena gracilis*)；*H. marinus*，海洋氢弧菌 (*Hydrogen ovibrio marinus*)；*M. acetivorans*，噬乙酸甲烷八叠球菌 (*Methanosarcina acetivorans*)；*N. tabacum*，*Nicotiana tabacum*；*O. luteus*，*Olisthodiscus luteus*；*P. hollandica*，荷兰原绿丝蓝细菌 (*Prochlorothrix hollandica*)；*S. oleracea*，*Spinacia oleracae*；*T. kodakaraenis*，超好热始原菌 (*Thermococcus kodakaraensis*)；*T. denitrifcans*，脱氮硫杆菌 (*Thiobacillus denitrificans*)；*X. flavus*，黄色黄杆菌 (*Xanthobacter flavus*)。

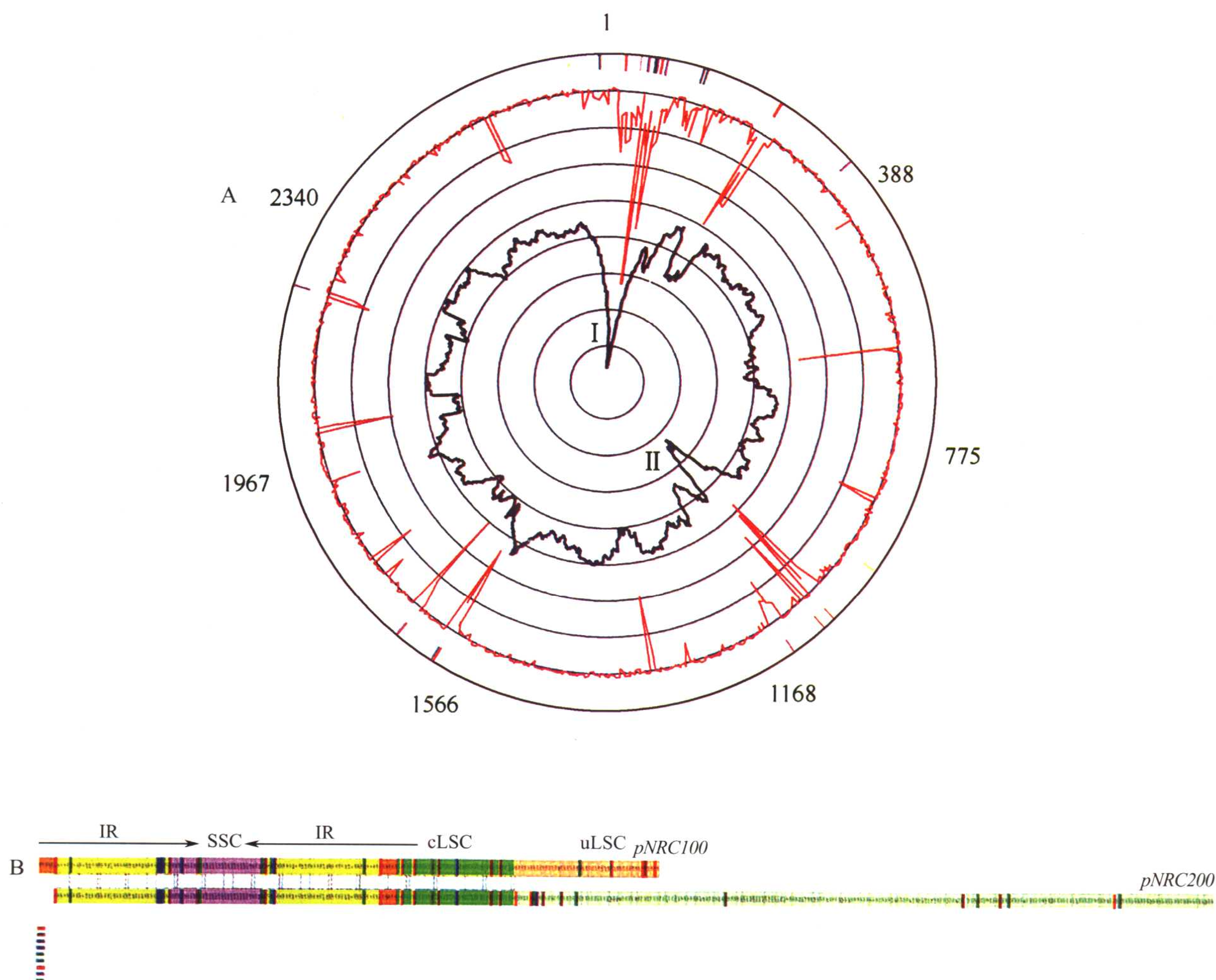


图 21-2 A. 盐杆菌 (*Halobaterium*) NRC1 大染色体环状图谱。B. pNRC100 和 pNRC200 复制子的线状遗传图谱比对。A. 大染色体环状图包含 IS 元件的位置 (外圈)，x 平方分析 (红线)，可读框 G+C 组成 (黑线)。与最外圈相连彩色线段表示染色体 IS 位置 (ISH1，浅褐色；ISH2，紫色；ISH3，绿色；ISH4，黄色；ISH6，粉色；ISH8，蓝色；ISH10，红色)。罗马数字 I、II 表示富 AT 岛。B. 以线状形式描述环状复制子，基因和 IS 元件用块状表示。两个复制子包含 145 428 bp 一致区和 45 918 bp (pNRC100) 或 219 997 bp (pNRC200) 的特有 DNA [3,4]。33kb 到 39kb 的反向重复用黄色 (在所有拷贝中保守) 和橙色 (在一些拷贝中保守) 表示；小单拷贝区用紫色表示；共同单个大拷贝区用亮绿色表示；独一无二的单个大拷贝区用棕褐色 (pNRC100) 和淡绿色 (pNRC200) 表示。IS 元件用暗橙色 (ISH2)、棕色 (ISH3)、靛蓝色 (ISH5)、蓝色 (ISH7)、暗绿色 (ISH8)、皂蓝 (ISH9)、蓝灰色 (ISH11)。两个 pNRC 复制子含 69 个 IS 元件 (44 个是独特的)，其中 29 个在 pNRC100 上、40 个在 pNRC200 上；有 6 个因子是反向重复 (在 pNRC100 和 pNRC200 中都重复 2 次)，在 pNRC100 和 pNRC200 中的 SSC 区中都有 4 个因子；在 pNRC100 和 pNRC200 中的共同单个大拷贝区都含 7 个因子；在独特单个大拷贝区含 23 个因子，其中 6 个在 pNRC100 上，17 个在 pNRC200 上。(图 2A 经冷泉港实验室出版社许可复制，文献[11])

OxIT formate oxalate
LctP lactate
RbsAC ribose
UgpABCE glycerol 3-phosphate
H^+ Respiratory Chain
ATP synthase ATP H^+ ADP_i P_i
BR H^+ HR hv Ret Ret Cl^-
SRI/HtrI SRII/HtrII(Ser) hv Ret Ret
Htr X (HemAT)(O_2)
Htr IX,XII,XIII XVII,XVIII(?)
CheAW CheY Flagellar Motor
Htr IV-VII
Htr XIV-XVI (?)
HtrVIII (O_2)
HtrIII(BasT) (Cys,Ile,Leu, Met,Val)
PotABCD spermidine putrescine
DppABCDF dip eptides
AppBCF oligopeptides
PutP proline
Cat cationic amino acids
amino acid ABC transporter amino acids
SecDEFY
Nutrient Uptake
Energy Production
Signal Transduction
Semi-Phosphorylated Entner-Doudoroff
glucose → gluconate →* 2-keto-3-deoxygluconate
2-keto-3-deoxy-6P-gluconate
gly ceraldehyde 3-P
pyruvate
pyruv ate
acetyl Co-A
acetate
Fatty Acid Oxidation
DNA REPAIR
Photoreactivation
Excision repair
Mismatch repair
Recombinational repair
sensor kinases
response regulators
sens or kinases
GV GV GV
aspartate
Citric Acid Cycle
arginine
glutamate
Protein Translocation
7S RNA
SRP54
SRP19
SRα
Neurosporene<Phytoene<Mevalonate
Retinal<Lycopene>Bacterioruberin
Anion Transport
Cl^- chloride channel
SO_4^{2-} CysAT
PO_4^{3-} PstABC
PO_4^{3-} phosphate permease
AsO_4^{3-} ArsC AsO_2^{1-} ArsAB AsO_2^{1-}
Cation Transport
K^+ TrkAH
K^+ KdpABC
H^+ Na^+ NhaC
Cd^{2+} ZntX
Cd^{2+} Cd efflux ATPase
Co^{2+} CbiNOQ
Cu^{2+} NosFY
Fe^{3+} iron permease
Fe^{3+} HemUV
Zn^{2+} ZurMA

图 21-4 盐杆菌（*Halobacterium*）NRC-1 基因组整体图。图中描绘了能量产生、营养吸收、膜组装、阴阳离子转运以及信号传导等几方面的内容。展示了通过由呼吸链的质子运输、由细菌视紫红质（BR，紫色卵形）光驱动的质子泵、由卤紫质（HR，蓝色卵形）进行的氯离子运输等化学渗透耦合（chemiosmotic coupling）产生 ATP 的过程。在下方，展示了半磷酸化 Entner. Doudorff 途径、脂肪酸氧化和三羧酸循环。星号标出的是尚未鉴定的酶。基因组编码了各种各样的营养吸收系统（黄色或棕色标示的结构），如甘油 -3- 磷酸运转蛋白（UgpABCE）和糖 ABC 运转蛋白（RbsAC）、乳酸盐运转蛋白（LctP）、甲酸盐草酸盐反向运转蛋白 (OxiT)、亚精胺与丁二胺 ABC 运转蛋白 (PotABCD)，氨基酸 (PutP, Cat)和二肽运转蛋白 (DppABCDF)。可能存在以通用 ABC 运转蛋白为代表的其他氨基酸吸收系统。图中黑色所示是蛋白移位器的几个组分 (SecDEFY、SRP19、SRP54、SR α)。图中标示出类胡萝卜素和视黄醛 (Retinal, Ret) 的生物合成。绿色所示是下列几种阳离子运输：K^+ (TrkAH 与 KdpABC)、Na^+ (NhaC)、Cd^{2+} (ZntX 与 Cd 外排 ATP 酶)、Co^{2+} (CbiNOQ)、Cu^{2+} (NosFY)、Fe^{3+} (铁离子透性酶与 HemUV) 和 Zn^{2+} (ZurMA)。红色所示是下列几种阴离子运输：SO4^{2-} (CysAT)、PO_4^{3-} (PtABC 与磷酸盐透性酶)、Cl^- (氯离子通道) 和砷酸盐 (ArsABC)。图中标示出由光受体与信号传导组分组成的复杂系统，包括两个感受器 (蓝色的是 SRI、橙色的是 SRII)、17 个响应光 (hv)、O_2 或氨基酸传导蛋白 (HtrIHtrX、HtrXII、HtrXVIII)。运动性信号经过 CheAW 和 CheY 向鞭毛马达的传递，如箭头所示。鞭毛绘成波浪线。图中列举了一个感受激酶［膜结合的 (白色菱形)或胞质的］和响应调节蛋白的例子。细胞内还绘出了气囊 (GV，白色卵形) 和 DNA 修复系统。

译者名单(按姓氏笔画为序)

王清锋 博士、研究助理 Section of Molecular Genetics and Microbiology, University of Texas Austin, USA. Email: qingfengw@hotmail.com

冯丽萍 博士后 Duke University. Email:jhwu74@yahoo.com

伍建宏 博士后 Duke University. Email: jhwu74@yahoo.com

刘子铎 博士、教授 武汉华中农业大学农业微生物学国家重点实验室。Email:lzd@mail.hzau.edu.cn

刘作易 博士、教授 贵州大学农学院,贵州省农业科学院。Email:liuzuoyi@yahoo.com.cn

刘明秋 博士 上海复旦大学生命科学学院。Email: liumq@fudan.edu.cn

刘 斌 博士生 Auburn University. Email:liubin1@auburn.edu

刘 超 博士生 武汉华中农业大学生命科学技术学院。Email: liuchao2003@webmail.hzau.edu.cn

吕颂雅 博士、副教授 武汉大学生命科学学院。Email:lusy68@sina.com.cn

孙 明 博士、教授 武汉华中农业大学农业微生物学国家重点实验室。Email:m98sun@mail.hzau.edu.cn

朱晨光 博士 上海交通大学生命科学技术学院。Email: zhuchenguang@sjtu.edu.cn

江 昊 博士 上海中科伍佰豪生物工程有限公司。Email: jhao928@yahoo.com

许朝晖 博士后 University of California, Los Angeles. Email:zhaohuixu2000@hotmail.com

吴天福 博士后 Southwestern Medical Center. Email: wutianfu@hotmail.com

张利莉 博士、教授 新疆塔里木大学。Email: zhang63lyly@yahoo.com.cn

张 琼 博士 生物芯片北京国家工程研究中心。Email:Cathyzhang@Capitalbio.com

汪世山 博士生 University of Massachusetts. Email:shishanw@foodsci.umass.edu

邵宗泽 博士、研究员 厦门国家海洋局第三海洋研究所。Email:shaozz@163.com

周世力 博士、副教授 武汉江汉大学生命科学学院。Email: z4lzjw@sohu.com

欧阳立明 博士 上海华东理工大学。Email:ouyanglm@eyou.com

徐进平 博士、副教授 武汉大学生命科学学院。Email: jpxu6077@163.com

喻子牛 教授 武汉华中农业大学农业微生物学国家重点实验室。Email:yz41@mail.hzau.edu.cn

喻晓辉 博士生 Clemson University. Email: xiaohuiyu69@hotmail.com

程 萍 博士、研究员 广东珠海市农业科学研究中心。Email:nkpcheng@163.com

译者的话

我室在承担国家863、973、国家自然科学基金等项目中，为了将功能基因方面的研究进行得更深入，想方设法对苏云金芽孢杆菌YBT-1520进行全序列测定，以期在这方面的研究能持续向更高水平发展。基因组学是一个崭新领域，在学习—工作—学习—工作—再学习—再工作的过程中，研究者们不得不翻阅国内外基因组学方面的大量文献。我作为美国微生物学会（AMS）二十多年的老会员，在收到寄来的书目中，看到Humana Press出版了*Microbial Genomes*，读了简介和目录后，像细菌对数生长需要营养那样，我迫不及待地打电话让我儿子从美国订购，很快便拿到手。看后得知，该书的主编是TIGR公司总裁C. M. Fraser博士等人，并由TIGR董事长作序，撰写人都是微生物基因组学方面的开拓者和“首次吃螃蟹的人”，撰写内容多为现身说法，这样的书应当视为精品。

从本书中读者可以了解到微生物基因组学的发展历史，虽是那么短暂，却蕴藏着微生物基因组学的深远意义、曲折历程、基础理论、研究方法、应用基础、主要焦点、发展方向、闪亮未来等丰富内涵，这正是我们当前研究所迫切需要的知识。由于科学技术的飞速发展，我翻了几遍原著却有很多内容一知半解，于是，联想起现在工作在生命科学第一线我的研究生们，他们富有朝气、思维敏捷、易于接受新知识并都富于工作经验，与他们通报翻译此书的想法后，得到他们的赞同和鼓励。原本要他们自己独立组织翻译，可他们非要我参与，最终确定由许朝晖博士牵头，我组织这些博士、教授们，他们在百忙之中接受了分配的任务。翻译稿全部交许朝晖博士校对，有些章节又反馈给译者核准，最后交给我根据汉语习惯，未死抠英语语法在文字上润色。清样由吴丹丹同学与译稿校对，刘超与原稿逐句校对。

感谢科技部中国生物工程中心给予的项目支持，感谢科学出版社及时申请到翻译版权，感谢昔日的研究生们给我今日再学习的机会。我在反复、逐字逐句修订译稿时，领会到我的导师、微生物学家陈华癸院士告诫“翻译好一本书不比写书容易”的忠言。我深信，本书中文版的出版不仅对微生物学，还将对其他相关学科的科教人员和研究生们，在面临基因组时代的挑战中有所裨益。

实话已实说，书中的谬误之处，敬请明眼人批评指正，以便再次印刷时更正。

喻子牛

2005年2月于武昌狮子山

华中农业大学生命科学技术学院

农业微生物学国家重点实验室

微生物农药国家工程研究中心

序

近10年前，即1995年，基因组研究所（TIGR）利用鸟枪法第一次进行了流感嗜血菌（*Haemophilus influenzae*）的全基因组测序尝试[1]。这次实验成功地引导科学进入了一个新时期。截至2004年，有150多种微生物基因组测序已经完成，还有上百种正在测序，微生物基因组学领域现已日趋成熟。微生物基因组计划已经从需要整个实验室或主要团队参与，并耗时长达几年到十几年的庞大工程，到如今研究生或博士后都能掌握，且只需几个星期到几个月，通过用基因组测序相关仪器设备即可完成。在早期，需要十几年才能完成微生物基因组测序，而基因组研究所对流感嗜血菌基因组的测序仅耗时4个月，这已经是大大缩短了。J. Craig Venter科学基金会联合技术中心（JTC）是从事快速DNA测序的机构，它服务于基因组研究所、生物能量替代研究所（IBEA）和基因组学发展中心（TCAG）。在那里，流感嗜血菌全基因组测序只需该中心一天序列测定工作量的25%，即在4小时内就可完成。在未来10年里，DNA测序技术将持续以对数形式提高，届时技术成熟的DNA测序中心每天可处理成千上万的基因组测序，即使是小研究室每天也可测若干个微生物基因组序列，并可作为标准解析步骤的一部分。毫不夸张地说，基因组技术在环境中应用，能在某一地域发现上百万的新基因[2]。我认为，这一新进展和不断扩大的微生物基因组测序工程，将带领我们更贴切地了解构成地球基因库的上百亿基因。

展望未来，微生物基因组学的主要挑战已越来越清晰地展现在我们面前，即基因组结构、进化和生物学。《微生物基因组》由Claire Fraser、Timothy Read和Karen Nelson主编，该书不仅将现代微生物基因组学开创以来做出主要贡献的科研人员汇于本书，而且还从基因组测序角度，对我们了解微生物新陈代谢和进化所起的作用提出了整体的看法，本书还适当地列出一些对基因组测序和分析中各阶段非常重要的生物信息工具，如果科研机构希望处理大量的基因和基因组数据，那么生物信息学和计算方法还需要得到相当重要的扩展。

我所称的基因组时代，即流感嗜血菌全基因组测序以后的时期，导致了药品和疫苗发展、微生物法医学、工业化学、生物治理等诸多方面的实际应用、新发现和新思路。就在几年前，那些听起来像科幻小说的想法和观念，如今已开始进行严密的实验，例如，美国能量部门希望将希瓦菌氏菌（*Shewanella*）基因组[3]中编码金属代谢途径基因，与耐辐射异常球菌（*Deinococcus radiodurans*）[4]基因组中抗辐射的基因结合起来，得到既能抗辐射又能分解铀的菌株；其他一些政府机构正试图利用不同微生物菌落分离株的多态性，来追踪与生物恐怖有关微生物的实验室来源，以及地域来源和个体活动；一些成功的公司正利用微生物基因组作为生产新型工业酶和化合物的来源，这一领域很可能随着合成基因组学的兴起而爆炸式地增长[5]。固然挑战很多，但微生物基因组学影响我们生活的机会将更多。《微生物基因组》一书能提供许多信息，在未来高速发展的微生物基因组学领域这些信息将起巨大作用。

J. Craig Venter, PhD

（喻子牛　译）

参考文献

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
2. Venter JC, Remington K, Heidelberg JF, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science in press.
3. Heidelberg JF, Paulsen IT, Nelson KE, et al. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. Nat. Biotechnol 2002; 20:1093–1094.
4. White O, Eisen JA, Heidelberg JF, et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 1999; 286:1571–1577.
5. Smith HO, Hutchison CA III, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci USA 2003; 100:15440–15445.

前言

在过去10年中，微生物学领域已转向全基因组序列研究，其中，位于马里兰州Rockville的基因组研究所（TIGR）在这一领域起了很大的作用，该所的科学家们公布了最早测得的三种微生物全基因组序列——流感嗜血菌（*Haemophilus influenzae*）、生殖道支原体（*Mycoplasma genitalium*）和詹氏甲烷球菌（*Methanococcus jannaschii*）。截至2004年1月，在已完成测序并公布的150多个基因组序列中，他们所公布的基因组序列有40多个。因此，借着这个令人鼓舞的机会，Humana出版公司计划出一本关于该领域研究进展的书，各章作者中很多是我们基因组研究所的同事，也有来自世界各地的一些专家。

微生物基因组学是一个非常大的领域，当选择每一章的题目时，我们意识到，要么涵盖较大范围而对相关知识仅作初步介绍，要么把我们认为非常重要的一些问题深入讨论，我们选择了前者。我们希望《微生物基因组》一书能描述更广的内容，能使初涉这一领域的读者和具有专业知识的人都对此书感兴趣。作为主编，我们发现一些很有趣的事情，尽管题目各不相同，某些重复出现的主题却贯穿全书，如基因水平转移、比较基因组分析的重要性和基于微阵列的基因表达分析。

《微生物基因组》分为6个主要部分，为让读者更好地了解这一领域的发展，分出一部分介绍微生物基因组学历史是十分必要的，所以主编委托Hamilton Smith准备这一介绍性文章。“作为基因组学工具的生物信息学”部分，提供了用于基因组学及其应用的最常用的计算工具。“核心功能”部分介绍了每个微生物基因组都包括的微生物代谢、运输和细胞循环等过程。“微生物基因组的进化”中，安排了一系列章节，目的是告诉大家，基因组学如何重建微生物世界的历史和动态性。在“微生物基因组的调查”中，组织了一系列章节并有选择性地论述了一些生物类群。尽管没有涵盖每种微生物基因组（而另一些种在多个章节中被描述），但我们的目的是提供一些从基因组研究可以获得的生物信息。最后是“基因组数据库的应用”，这部分描述如何用基因组序列研究微生物中最重要的问题。值得一提的是，在本书撰写期间，即2002年底到2003年初，许多尚未完成的基因组测序现已完成，感兴趣的读者可以访问像基因组研究所数据库（www.tigr.org/tdb/mdb/mdbcomplete.html）这样的网站，以便得到最新的信息。

如果没有Trina Eacho女士的帮助，这本书不可能成功出版，她利用很多周末时间组织各个章节并形成初稿，主编们对她的诸多贡献非常感激。

当然，此书难免有疏漏之处，而且它也不能涵盖所有的重要方面，我们希望著者们的集思广益及各章节的内容能给诸位读者以启发。微生物基因组学包括了进化与种群生物学、基因表达分析、蛋白质组学以及很多与DNA序列相关的研究。毫不夸张地说，微生物基因组学跨越了微生物学科本身的广度和深度。

Claire M. Fraser, PhD
Timothy D. Read, PhD
Karen E. Nelson, PhD
（喻子牛　译）

编著者

SIV G. E. ANDERSSON, PhD • *Department of Molecular Evolution, Uppsala University, Uppsala, Sweden*
STEPHEN D. BENTLEY, PhD • *The Pathogen Group, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
SVEND BIRKELUND, MD, PhD, DMSc • *Department of Medical Microbiology and Immunology, University of Aarhus, Aarhus C, Denmark*
FIONA S. L. BRINKMAN, PhD • *Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada*
ROLAND BROSCH, PhD • *Unite de Genetique Moleculaire Bacterienne, Institut Pasteur, Paris, France*
C. ROBIN BUELL, PhD • *Department of Plant Genomics, The Institute for Genomic Research, Rockville, MD*
GUNNA CHRISTIANSEN, MD, DMSc • *Department of Medical Microbiology and Immunology, University of Aarhus, Aarhus C, Denmark*
FREDERICK M. COHAN, PhD • *Biology Department, Wesleyan University, Middletown, CT*
STEWART T. COLE, PhD • *Unite de Genetique Moleculaire Bacterienne, Institut Pasteur, Paris, France*
SHILADITYA DASSARMA, PhD • *Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, MD*
ARTHUR L. DELCHER, PhD • *Department of Bioinformatics, The Institute for Genomic Research, Rockville, MD*
EDWARD F. DELONG, PhD • *Monterey Bay Aquarium Research Institute, Moss Landing, CA*
ROBERT A. FELDMAN, PhD • *SymBio Corporation, Menlo Park, CA*
DERRICK E. FOUTS, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
JOANNA L. FUEYO, PhD • *IBM Life Science Solutions Consulting, Cambridge, MA*
MALCOLM J. GARDNER, PhD • *Department of Parasite Genomics, The Institute for Genomic Research, Rockville, MD*
STEVEN R. GILL, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
J. PETER GOGARTEN, PhD • *Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT*
STEPHEN V. GORDON, PhD • *Veterinary Laboratories Agency, Addlestone, Surrey, UK*
GUIDO GRANDI, PhD • *Chiron Corporation, Siena, Italy*

Mark E. Hance • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Thomas E. Hanson, PhD • *Department of Microbiology and the Plant Molecular Biology/Biotechnology Program, The Ohio State University, Columbus, OH*
David A. Hopwood, PhD • *Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Colney, Norwich, UK*
Katherine H. Kang• *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Robert M. Kelly, PhD • *Department of Chemical Engineering, North Carolina State University, Raleigh, NC*
Michael T. Laub, PhD • *Bauer Center for Genomics Research, Harvard University, Cambridge, MA*
Vega Masignani, PhD • *Chiron Corporation, Siena, Italy*
Harley H. McAdams, PhD • *Department of Developmental Biology, Stanford University, Palo Alto, CA*
Garry S. A. Myers, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Karen E. Nelson, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Lorraine Olendzenski • *Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT*
Julian Parkhill, PhD • *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
Ian T. Paulsen, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Mariagrazia Pizza, PhD • *Chiron Corporation, Siena, Italy*
Rino Rappuoli, PhD • *Chiron Corporation, Siena, Italy*
Jacques Ravel, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Timothy D. Read, PhD • *Biological Defense Research Directorate, Naval Medical Research Center, Silver Spring, MD*
Qinghu Ren, PhD • *Department of Microbial Genomics, The Institute for Genomic Research, Rockville, MD*
Frank T. Robb, PhD • *Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, MD*
Carsten Rosenow, PhD • *Affymetrix, Santa Clara, CA*
Steven L. Salzberg, PhD • *The Institute for Genomic Research, Rockville, MD and Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD*
Lucy Shapiro, PhD • *Department of Developmental Biology, Stanford University, Palo Alto, CA*
Allan C. Shaw, PhD • *Department of Medical Microbiology and Immunology, University of Aarhus, Aarhus C, Denmark and Novo Nordisk Research and Development Center China, Beijing, China*

Keith R. Shockley, PhD • *Department of Chemical Engineering, North Carolina State University, Raleigh, NC*
Hamilton O. Smith, PhD • *Institute for Biological Energy Alternatives, Rockville, MD*
F. Robert Tabita, PhD • *Department of Microbiology and the Plant Molecular Biology/Biotechnology Program, The Ohio State University, Columbus, OH*
John L. Telford, PhD • *Chiron Corporation, Siena, Italy*
Nicholas R. Thomson, PhD • *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
Brian Tjaden, PhD • *Department of Computer Science, Wellesley College, Wellesley, MA*
Brian B. Vandahl, MSc • *Department of Medical Microbiology and Immunology, University of Aarhus, and Loke Diagnostics ApS, Aarhus C, Denmark*
J. Craig Venter, PhD • *The Center for the Advancement of Genomics, Rockville, MD*
Lawrence P. Wackett, PhD • *Department of Biochemistry, Molecular Biology an Biophysics, University of Minnesota, St. Paul, MN*
Owen White, PhD • *The Institute for Genomic Research, Rockville, MD*
Olga Zhaxybayeva • *Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT*

目录

第一部分：引言

1

微生物基因组学的历史

Hamilton O. Smith

引言

基因组学*是一门对生命有机体全基因组序列进行分析和比较的新兴学科。基因组序列为我们提供了有机体的最基本信息，序列中的基因和调控位点就是该有机体的“零部件”和“运行指令”，同时它还为该有机体提供进化依据，序列就自然而然地成为研究诸多新物种的出发点。基因组学是20世纪医学和生物学飞跃发展中最激动人心的成果之一，基因组学兴起于20世纪最后10年，并为21世纪的医学和生物学打下了坚实的基础。

我的研究兴趣集中在微生物基因组学。它的兴起与人类基因组计划（Human Genome Project，HGP）密不可分，毫无疑问，没有人类基因组计划，第一个微生物基因组的序列测定就会大大推迟。人类基因组计划前5年中就提议对两种模式微生物，大肠杆菌（*Escherichia coli*）和酿酒酵母（*Saccharomyces cerevisiae*）进行测序。这是最早并始于1990年左右的微生物基因组项目。

因此，有必要简要地回顾人类基因组计划，并顺便介绍一下该计划之前的知识水平。基因组学和人类基因组计划并不是空穴来风，贯穿于整个20世纪生物学领域的研究和发展，为该计划提供了基本的知识和技术手段。如果不知道脱氧核糖核酸（deoxyribonucleic acid，DNA）的化学本质，就不可能发明DNA测序技术，也就不会有现代基因组学。因此，本章将为现代基因组学介绍一些铺路搭桥的事件和发现，以及一些具有里程碑意义的测序项目，这些项目使得微生物基因组学发展成为一个令人瞩目的新兴科研领域。

对历史事件的回顾主要集中在微生物基因组的测序。测序后紧接着分析序列，这是目前正在飞速发展的方面，作为历史来描述不太适合。此外，本书还有相当篇幅介绍基因组分析以及从基因组序列中获得生物信息的各种方法。

由于本人阅历所限，对相关历史事件难免有所偏颇，自然会对一些我亲眼目睹和记忆的往事详加叙述并着重强调。因此，对历史回顾不可能十分完善，也不可能对所有参与者的贡献给予非常公正和均衡的肯定，为此我要向大家致歉。

---

* 基因组学是个新名词。1986年，杰克逊实验室（Jackson Laboratories）的Tom Roderick提议用它来命名旨在研究全基因组序列及与之相关高通量（high-throughput）技术的新兴学科[10]。1987年，Victor McKusick创办了新杂志《基因组学》。

基因、染色体、基因组、噬菌体、细菌和 DNA

基因组（genome）一词是 1920 年由 Winkler 引入学术界的，它由基因（GENe）和染色体（chromosOME）两个词组合而成[2]，代表完整的单套染色体和基因。当时对基因的认识很肤浅，只知道它们是决定动植物的可视或可测量性状的遗传单位，其化学本质完全是个未知数，对植物、果蝇和人类的经典遗传学研究还不能回答这个问题。但在 20 世纪 30～50 年代，对细菌和噬菌体等简单生物的研究导致了染色体、基因和 DNA 本质的重大发现。

有趣的是，由于细菌没有细胞核和高等生物那样的染色体，因此，最初认为它好像能以非遗传学机制快速适应环境。1943 年 Luria 和 Delbruck 用波动实验（fluctuation analysis）证明，适应只是对细菌群体中已经存在的突变株进行遗传选择[3]，细菌有同其他生命一样的遗传特性和基因。1944 年，Avery 及她在洛克菲勒研究所（Rockefeller Institute）的同事们[4]证明，DNA 是细菌转化实验中遗传物质的载体。

1946 年 Lederberg 和 Tatum[5]发现，将大肠杆菌两株不同突变株混合培养可以得到新的重组菌。在其后近 10 年中，几个实验室的研究证实只有部分菌株是可育的，在菌株交配时，雄性菌按照一定的顺序将部分染色体转移给雌性菌[6]。根据遗传标记进入雌性菌的顺序就可以画出遗传图谱，随着越来越多的标记被定位，该图谱最终呈现为环状。

与此同时，1953 年 Watson 和 Crick 根据 X 射线衍射照片推断出 DNA 的结构是由两条反向平行的单链组成的双螺旋结构[7]。他们推测双螺旋上的碱基序列是遗传信息的载体，并进一步预言 DNA 是通过半保留方式复制，拷贝互补的母链就可以产生新的子链。

1963 年，Cairns 用氚标记的胸腺嘧啶核苷标记大肠杆菌的 DNA[8]，经放射自显影技术证明，大肠杆菌的基因组是一个以半保留方式复制的单链环状 DNA 分子。

这些实验毫无疑问地证明，细菌是以 DNA 染色体为遗传物质的有机体。

直到 1990 年，也就是基因组学时代即将到来之时，大肠杆菌[9]、鼠伤寒沙门氏菌（*Salmonella typhimurium*）[10]和枯草芽孢杆菌（*Bacillus subtilis*）[11]的遗传图谱已相当详细，包括成百上千个定位基因，靠这些图谱几乎可以诞生低精确度的比较基因组学了。当时，谁也没料到利用重组方法构建遗传图谱的时代将要结束了，20 世纪 90 年代中期对这些细菌的全序列测定，在很大程度上取代了前几十年在实验室构建图谱所取得的成就。

DNA 测序技术的发明

没有 DNA 测序，就没有真正意义上的微生物基因组学。或许你可以说基因组学是以有机体的基因组为研究对象，因此它伴随着遗传图谱和限制性酶切图谱的产生而产生。但是，要总览单倍染色体组所包含的所有遗传信息（即待研究的有机体所有基因），并将它们与其他有机体进行比较（这正是基因组学的精髓），仅凭那些用“粗放”手段

搜集的相当有限的数据是远远不够的，只有知道了组成基因组的所有 DNA 序列，才能达到上述要求。因此，基因组学只有在 DNA 测序发明之后才能成为一门羽翼丰满的学科。

1975 年 6 月，Gordon 会议在新英格兰召开，会上发生了一件大事，两个相互独立的研究组激动地宣布他们各自发明了 DNA 测序技术，尽管两种方法都不够完善，人们却已经意识到 DNA 测序的新时代就要到来。

两种方法的共同点是将 DNA 5′端固定，其区别在于碱基特异性切割或终止方式不同。在不同碱基处终止 DNA 片段，可以通过聚丙烯酰胺凝胶电泳一条条地分开，Maxam[12]采用的是碱基特异性的化学消化法，这样 A 和 G 就分布在一条泳道上，C 和 T 则分布在另一条泳道上，根据电泳带的浓度可以区分出 A 和 G，以及 C 和 T。1976 年初，Maxam 和 Gilbert 的方法进一步完善为四泳道法，不同碱基分布在不同的泳道上，直到 1980 年他们才发表了这个新方法的详细过程。Sanger 研究组报道了八泳道酶法测序，也就是所谓的加减法，该法要用 DNA 聚合酶。1977 年，Sanger 及其同事改进了该法，用四种双脱氧核苷酸进行链终止反应[13]。1978 年，Sanger 和 Coulson 引进了超薄胶以后[14]，便可从一块胶上读出几百个碱基。

自动测序仪的发明

从 20 世纪 70 年代末到 80 年代初，采用两种方法中的任何一种，在生物学实验室经常对长达几千个碱基的 DNA 片段进行测序。但是，Maxam 和 Gilbert 的化学法不如 Sanger 及其同事的酶法简便，很快就被淘汰了。1977 年，Sanger 研究组完成了第一个全基因组——ϕX174 噬菌体基因组（5386bp）测序[15]，1982 年，该室又完成了 λ 噬菌体基因组（48 502bp）测序，这是当时最大的测序工程[16]。测序是个劳动密集型的工作，它要先用同位素标记核苷酸，再放射自显影，然后花大量精力读胶，这都不可避免地容易出错，在随后的几年中却大大改观了，测序将变得更快更自动化。

1985 年，在位于帕萨迪纳（Pasadena）的加州理工学院（California Institute of Technology）工作的 Hood 和 Smith 向人们展示了用四种荧光染料标记 DNA 的方法，这样就可以用自动激光仪阅读测序胶片了。1986 年 6 月，他们宣布第一台自动 DNA 测序仪诞生了[17]，1987 年底，应用生物系统公司（Applied Biosystem Inc.）采用 Hood 的技术开发了第一台上市的自动测序仪。每台仪器每天可以测 1～2 万个碱基粗序列（raw sequence）。而现在的毛细管测序仪，如 ABI Prism 3700（应用生物系统公司，加州福斯特市，Foster City），每台仪器每天可以测出 50 万个碱基粗序列。

人类基因组计划

微生物基因组学的历史与美国能源部（US Department of Energy，DOE）和人类基因组计划有不可分割的联系，如果没有人类基因组计划，微生物基因组学就不可能有现在这么先进。人类基因组计划抓住了人们的想像力，给我们提供了经费，并能前所未有地鼓舞和推动科学家们去开发工具和寻找策略。

1986 年 3 月，美国能源部健康与环境研究办公室（Office of Health and Environ-

mental Research）的 Charles DeLisi 和 David Smith，在新墨西哥州圣菲市（Sante Fe, New Mexico）主持召开了一次会议，与会的 30 多名科学家讨论了测定人类基因组的可行性。我有幸参加了这次会议，清楚记得和很惊讶地看到与会者们几乎一致对该项目表现出极大的热情。尽管当时的测序技术还没有现在这么先进（自动测序仪一年以后才上市），还不一定能切实可行地行使这一具有纪念意义的使命，议题却主要围绕策略和花费，大家讨论了各种策略，包括酵母人工染色体、噬菌体、黏粒图谱（cosmid map），随机鸟枪法测序（random shotgun sequencing）和 cDNA 等。大多数人主张用图谱，用大量酵母人工染色体和黏粒克隆来交叠覆盖人类基因组，然后再对单个克隆测序，依此估计，每完成一个碱基要花费一美元，整个项目需要 30 亿美元。

几乎同时，Dulbecco 在《科学》杂志[18]上发表了一篇评论，强力支持和提倡人类基因组计划。1986 年 9 月，为了启动人类基因组计划，DeLisi 从能源部拨款 530 万美元给该部的国家实验室做前期研究。1987 年，能源部成立了顾问委员会，该委员会建议在 7 年中拨款 10 亿美元用来建图谱和测序，并由能源部领导美国方面的工作。随着人们热情的高涨，美国国家研究委员会（National Research Council）也于 1988 年开始支持人类基因组计划，并呼吁每年投资 2 亿美元。

同年，美国国立卫生院（National Institute of Health, NIH）主任 James Wyngaarden 认为，人类基因组计划与健康有关，NIH 应该是主要参与者。事实上，NIH 虽然姗姗来迟，却从能源部手中抢走了领导权，James Watson 被任命为新成立的人类基因组研究办公室（Office of Human Genome Research）主任，启动项目所需的经费也拨了下来。为相互合作，能源部和 NIH 签署了谅解备忘录，由 NIH 领导负责人类基因组计划。

1990 年，能源部和 NIH 联合向美国国会提交了人类基因组计划的 15 年规划和 5 年研究计划。1990 年 10 月 1 日，人类基因组计划正式启动，这对微生物基因组学十分重要，在前 5 年计划中就提出对几种模式生物进行测序，其中包括研究最广泛的微生物：大肠杆菌和酵母。

大肠杆菌基因组：一个人的执着

1983 年，Blattner 首先提出测定大肠杆菌基因组[19]。大肠杆菌是当之无愧最重要的细菌，全世界几千个实验室都用它来研究生命体系的基本过程，以它作为重组 DNA 的载体。Blattner 是威斯康星大学麦迪逊分校（University of Wisconsin-Madison）的教授，他一生研究大肠杆菌和 λ 噬菌体，1977 年他构建的 Charon 噬菌体作为载体，操作安全并广为使用[20]。

1988 年前，Blattner 为大肠杆菌基因组测序构建了一套至少 15～20kb 交叠的 λ 噬菌体克隆。他从人类基因组计划中心得到一笔钱，1990 年开始测序，刚开始，他的策略是测定那些覆盖几百个 kb 的交叠克隆，在 1992～1995 年间，他们采用放射性标记，人工测定了总长 1.92 Mb 序列（基因组从 2 686 777 位点到 4 639 221 位点）。

1995 年，在全基因组鸟枪法被证明更为有效之后（见下文），对 Blattner 加快进度的压力越来越大，在申请继续测序经费时，他不得不在激烈的竞争中极力为自己辩护，并确定采用一种能够保证在 1 年内完成任务的新策略，最终他的申请被批准了。

新策略放弃了噬菌体克隆，而改用一种位点罕见的限制性内切核酸酶 I-SceI 酶，切

出的每个片段为250kb左右，它们覆盖了大部分剩余的基因组，然后，将这些片段用鸟枪法由自动测序仪测序，Blattner如期完成了任务，并于1997年9月发表了大肠杆菌的基因组序列[21]。这一期待已久的结果，无论是对大肠杆菌的研究者，还是对用该菌作分子遗传基本工具的广大研究者都具有十分重大的意义。

酵母基因组：国际合作的成功典范

酿酒酵母的测序是人类基因组计划前期项目中的另一个模式体系。它始于20世纪80年代末，完成于1996年[22]。酿酒酵母的基因组大小为12Mb，含有6000个基因，分布于16条染色体上，这是当时最大的微生物项目。酵母测序是现代分子生物学中最大最分散的实验[23]，包括美国、加拿大、欧洲和日本等国的100多个实验室的600多位科学家参与了该项目。在比利时Catholique de Louvain大学具有“杰出教授”荣誉的Andre Goffeau领导下，酵母学术圈是效率、合作和组织的典范，这种合作一直延续到测序后阶段，多个实验室共同揭示基因的功能和表达模式。

流感嗜血菌Rd：第一个完成的微生物基因组

流感嗜血菌从来没有被选作或被认为是模式生物，它的测序比大肠杆菌和酵母晚得多。令人惊奇的是，它又怎样成为第一个完成的微生物基因组呢？

费城宾夕法尼亚大学（University of Pennsylvania）的Sol Goodgal最先提议测序流感嗜血菌Rd。1988年他向HGP申请经费，打算用转座子随机插入法测序，由于人们对这一策略的可行性持怀疑态度，课题没有被批准。1993年夏，英国牛津Radcliffe医院的Richard Moxon到马里兰州巴尔的摩工作的约翰霍普金斯大学（John Hopkins）来访问。他长期从事流感嗜血菌血清型b毒力决定因子的研究，因而也提议测序流感嗜血菌。他曾是约翰霍普金斯大学的医学教授，与我一起共事多年，他建议我们两个实验室联合测序，我当时对他的想法并不“感冒”，曾反驳说该课题对我们两个专搞学术研究的实验室太大了，要想申请到经费几乎不可能，现在看来，我当时的判断是对的。

同年夏天，在西班牙Bilbao召开的基因组伦理会上（Genome Ethics），我有幸遇到了J. Craig Venter（图1），他邀请我加入基因组研究所（The Institute for Genome Research，TIGR）科学顾问委员会（Scientific Advisory Council）。Venter于1992年6月创办了TIGR，为的是用表达序列标签（expressed sequence tag，EST）快速识别人类基因，他创建的这个方法涉及到多种组织cDNA克隆的随机测序。

1993年9月召开了TIGR科学顾问委员会第一次会议，在讨论中我得知EST工作只要几个月就可以完成。当时，UIGR有30台373型ABI测序仪，每天可测四十万个碱基（图1）。不知为什么，我问Venter是否有兴趣测流感嗜血菌Rd的基因组？根据约翰霍普金斯实验室脉冲电泳结果判断，它的基因组大小约1.9Mb，而40%的G+C碱基组成也有利于测序，Venter表示了极大的兴趣，主动要我构建测序的文库，并在图谱上定位。

回到霍普金斯，我把这件事告诉了研究组其他人，出乎意料，他们都持怀疑态度。每个人都忙于自己的课题，建立克隆图谱至少要花1年，实验室又没有开展这项工作的经费，再申请经费也要好几个月。几个星期后，Venter问我文库建得怎样？我回答无

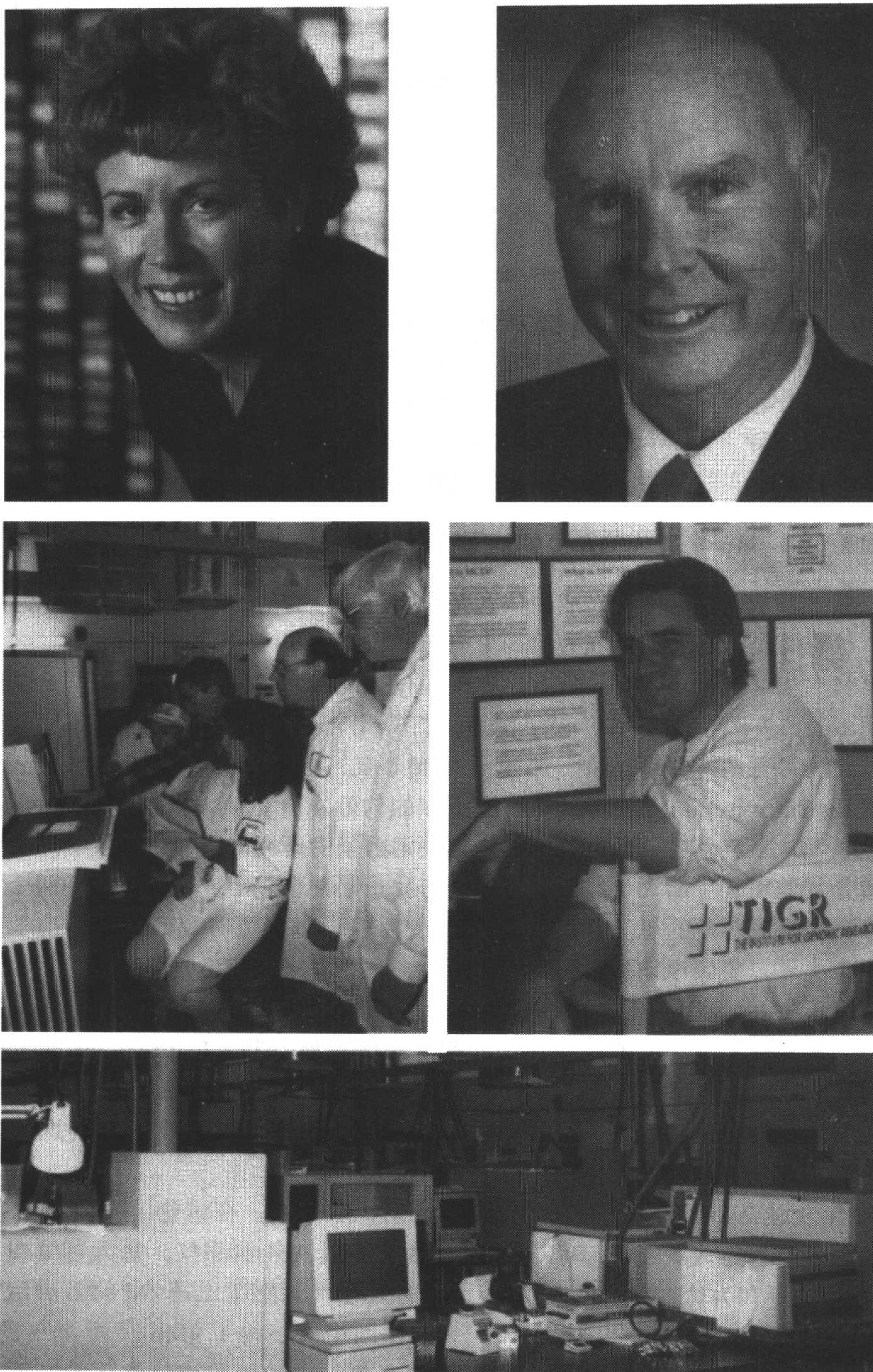


图 1　上左：Claire M. Fraser，基因组研究所（The Institute of Genome Research，TIGR）总裁。上右：J. Craig Venter，TIGR 创建者及董事会主席。中左：1994 年流感嗜血菌测序时 TIGR 研究组部分成员，前右，Hamilton Smith 站在 J. Craig Venter 旁边。Robert Fleischman 正指着电脑屏幕。中右：Owen White，TIGR 信息学主任。下：TIGR1995 年的测序实验室。

法像原先想像的那样构建和定位克隆，并建议应该考虑一条新途径。

我知道 TIGR 很擅长高通量（high-throughput）随机测序，为什么不构建一个短插入片段克隆文库，然后对它们进行随机测序，产生一批读序（或称标签，tag），最后再用计算机把序列组合到一起？TIGR 已经研发了一个程序，它能将成千上万个 EST 拼成一致 cDNA（consensus cDNA），这个 EST 策略完全可以用在细菌染色体的组合上。

1993 年 11 月，随机全基因组鸟枪法被正式介绍给 TIGR 的全体职员，并激发了大家极大的兴趣。于是，我们提交了一份 NIH 的资助申请，但是，因为随机鸟枪组合法的可行性问题而被否决了。尽管如此，Venter 完全相信这个策略会成功，决定动用 TIGR 筹来的经费开展这项工作，给这个课题开了绿灯。

1994 年初，霍普金斯构建了流感嗜血菌 Rd 1.8kb 片段的文库，然后由 TIGR 测序。TIGR 的科学家 Robert Fleischman 被任命为该课题组的组长，随机克隆文库于 1994 年 4 月开始测序，历时四个月完成。该序列“草图”有 24 000 个读序，平均每个 500bp，另外还有从 300 个左右大的片段（16 - 20kb）λ 噬菌体克隆得到的插入片段两个末端的读序。组合软件由 TIGR 的 Granger Sutton 设计，并在测序时被测试和完善。测序总量相当于把 1.8Mb 基因组的每个碱基读了 6 遍（用基因组学行话是 6 倍覆盖率），最初组合中有几十个短缺口，用其周围的序列设计引物进行聚合酶链反应就可把缺口补平，整个基因组测序和注释 1 年完成[24]。

朵美酒店（Dormy House）的专题研讨会及生殖道支原体（*Mycoplasma Genitalium*）

1995 年初，在 Venter 的鼓励和授权下，我和 Moxon 为了展示第一个细菌基因组序列，便组织了一次小型专题研讨会，那是一个有潜在历史意义的事件。由威康信托（Wellcome Trust）主持，1995 年 4 月在英格兰伍斯特郡（Worcestershire）的朵美酒店（Dormy House）举行。临近会期，威康信托中有人传言流感嗜血菌的序列并没有完成，Venter 不会交付或不允许公开数据。

当时，确实有正当理由怀疑 Venter 是否有权公开数据。TIGR 虽然是一家非营利性研究结构，但它却隶属于位于马里兰州罗克维尔（Rockville, MD）的人类基因组公司（Human Genome Science），它对 TIGR 的所有数据拥有所有权。在流感嗜血菌课题完成后的几周内，Venter 做了极大的努力，跟人类基因组公司总裁 William Haseltine 讨价还价，争取发表流感嗜血菌基因组的权利，争取终于得到回报，同意向会议交付流感嗜血菌的数据。但是，Venter 要求独立自主地发表 TIGR 研究数据的权利，跟 Haseltine 的这一冲突最终导致 TIGR 失去了三千五百万美元的捐助。

在流感嗜血菌的研究还处在最后收尾和注释时，Venter 就在考虑测序第二个基因组了，这就是他的风格，他急切地想展示鸟枪法的威力，同时也要证实流感嗜血菌的成功并不是偶然的。1995 年 2 月在一次午餐中，大家围绕哪个菌是第二个基因组测序的最佳选择进行了讨论，生殖道支原体是已知最小的细菌，其基因组大小估计不超过 600kb，显然是最佳选择。

大家要我给北卡大学（University of North Carolina）的 Clyde Hutchison 打电话，请他跟我们合作。Hutchison 研究生殖道支原体多年了，已经进行了约 350 个克隆的随机鸟枪法测序[25]，可能有所需的最小基因组，曾被邀请在朵美会议上作报告。我设法使他相信生殖道支原体的全基因组测序在会前就能结束，即在仅两个月之内完成。Hutchison 与同事们商量后很快答应与我们合作，并提供 10μg DNA 用来建文库。

在 Venter 妻子 Claire Fraser（图 1）领导下测序进展得很顺利，不到四周就完成了。基因组注释中发现了 470 个基因，该基因组在会上作为第二个基因组报道，因其基因组小而倍受关注。

全基因组鸟枪法掀起了微生物测序的高潮

在大肠杆菌和酵母基因组项目展开以后，微生物基因组学就逐渐成为一个完整的、被大家公认的研究领域。但是，直到 1995 年第一个具细胞形态的基因组，即流感嗜血菌 Rd 的基因组测序完成后[24]，它的地位才真正得以稳固。这项研究具有双重意义，首先它证明了生命有机体全套基因目录的价值；其次，证明了无论是否研究过的微生物都具有用全基因组鸟枪法进行基因组测序的潜力。

用全基因组鸟枪-装配法（whole-genome shotgun-and-assembly）进行微生物基因组测序的重要性也不能被低估。1995 年流感嗜血菌基因组发表后的 6 年里，微生物基因组测序掀起了极大的高潮，这在很大程度上归功于全基因组鸟枪法[24]。从环境中或从任何来源得到的未知或未曾研究过有机体的几微克基因组 DNA 都适合于测序。

一般情况下，需要构建一个小插入片段（2kb）文库和至少一个大插入片段（通常 10kb）文库，随机挑选足够数量的克隆，从插入片段的两端测序可以达到 6～10 倍覆盖率，就会给随后的序列组装提供极大的方便，从插入片段两端测得序列之间的距离是已知的，这种距离上的信息很有利于组装。

TIGR 仍在继续探索全基因组鸟枪法，截至 2002 年 9 月，他们已完成 21 种微生物基因组测序。

能源部与微生物基因组的起步

能源部 1994 年首先启动了微生物基因组计划（microbial genome project，MGP)，又一次在基因组学中发挥了先锋作用。他们的兴趣集中在非病原菌上，尤其是那些与环境、系统发育、商业或能源相关的微生物。早先他们与 TIGR 商讨时，同意为生殖道支原体提供资助，因为他们也对定义最小的基因组感兴趣。1995 年，在生殖道支原体项目取得巨大成功后，他们与 TIGR 和伊利诺斯大学（University of Illinois）签署了三年合作协议，测序詹氏甲烷球菌（*Methanococcus jannaschii*）和其他感兴趣的微生物。他们还与基因组治疗公司（Genome Therapeutics Corporation）（马萨诸塞州沃塞市，Waltham，MA)、犹他大学盐湖城分校（University of Utah in Salt Lake)、重组生物催化公司（位于加州圣迭戈 Recombinant BioCatalysis Inc. 即现在的达沃斯公司 Diversa Corp.）分别签署协议，测序热自养甲烷杆菌（*Methanobacterium thermoautotrophicum*)、

激烈火球菌（*Pyrococcus furiosus*），以及 *Aquifex aeolicus* VF5。*Aquifex* 是真细菌，其他三种是古生菌，因此，他们希望一次就得到第三域（domain）中三种古生菌的完整基因组。

古生菌早先被划分为细菌，但在 1977 年，伊利诺斯大学厄巴纳分校（Univeristy of Illinois in Urbana）的 Carl Woese，对两种产甲烷菌的 16S 核糖体 DNA 序列进行比较后发现，尽管两种菌亲缘关系接近，它们却与典型的原核细菌没有多少类似之处[26]。并建议把这些基本的生命形式叫古生菌[27]。直到 1987 年，已经有足够的生化和 16S 核糖核酸（ribonucleic acid，RNA）分类证据，使 Woese 及其同事们[28]提议在界（kingdom）之上设立一个全新的分类单元：域（domain），即把地球上的生命划分为三个域：古生菌（archaea）、细菌（bacteria）和真核生物（eukarya）。

詹氏甲烷球菌是第一个全序列测定的古生菌[29]。研究发现，它几乎三分之二的基因与以前发现的不同，有关转录、翻译和复制的基因与真核生物类似，而与生物合成和代谢有关的基因则与原核生物类似，这就很准确地把詹氏甲烷球菌在进化上划分到与细菌和真核生物都不同的古生菌中，从而巩固了 Woese 的假说。

能源部的微生物基因组计划仍在继续资助许多微生物测序项目，该计划在短短的 7 年中取得了非凡的成就。截至 2002 年 1 月，受资助者已经测序并发表了 6 种古生菌和 8 种真细菌的基因组，还有 10 种真细菌的测序已经完成，尚未发表，另有 14 种真细菌、2 种古生菌和 1 种真核生物已有测序草图。

日本与光合蓝细菌集胞藻（*Synechocystis*）PCC 6803

日本木更津 DNA 研究所（Kisarazu DNA Institute）本应成为第一个完成微生物基因组测序的机构。他们关于集胞藻菌株 PCC 6803 的杰出工作鲜为人知，1994 年发表了该菌基因组物理图谱[30]，1996 年完成并发表了 3.57Mb 基因组的测序和注释[31]，这是当时最大的基因组，并且比大肠杆菌和酵母菌的测序早一年完成，确实是件很了不起的工作。

国立过敏、传染病和人类病原体研究所（National Institute of Allergy and Infectious Disease and Human Pathogens）

国立过敏、传染病和人类病原体研究所起到了和能源部互补的作用。截至 2002 年 9 月，该所资助了 15 种重要人类病原菌的测序，正在资助或参与资助 44 种与人类健康相关的细菌、真菌和寄生虫的测序项目。

桑格研究所（Sanger Institute）和威康信托

英国亨克斯顿（Hinxton）的桑格研究所是美国之外最大的基因组中心。该所由威康信托资助，旨在英国建立一个人类基因组和其他基因组的重要图谱标记和测序中心。威康信托是世界上最大的慈善机构，拥有 150 亿英镑的资产，其宗旨是在促进动物和人

类健康的科学研究中起领导作用。在所长 John Sulston 的带领下，该所的近 600 名职员完成了人类基因组公共项目中近三分之一的测序任务。

桑格研究所专注于病原生物和模式生物，已经成为微生物基因组学的动力站。截至 2002 年 9 月，该所已发表了 7 种细菌基因组序列，还测完了 7 种菌的序列，另有 24 个项目处于不同进展阶段。桑格研究所正在测序的 5 种真菌是：裂殖酵母（*Schizosaccharomyces pombe*）、烟曲霉（*Aspergillus fumigatus*）、卡氏肺囊虫（*Pneumocystis carinii*）、白色念珠菌（*Candida albicans*）以及酿酒酵母，有的是独立测序，有的是参与测序。此外，他们还正在测序或参与测序 11 种原生动物病原体，包括刚完成的疟疾寄生虫的恶性疟原虫（*Plasmodium falciparum*）测序。

能源部基因组联合研究所（DOE Joint Genome Institute）

微生物测序的快速发展方兴未艾。加州核桃溪（Walnut Creek）的美国能源部基因组联合研究所，在过去几年内为微生物基因组测序做出了不可磨灭的贡献。他们已经完成了硝化自养菌欧洲亚硝化单胞菌（*Nitrosomonas europaea*）和 4 种光合菌的测序：蓝细菌中多种系群（polyphyletic group）的两个成员——海洋原绿球藻（*Prochlorococcus marinus*）菌株 MED4 和菌株 MIT9313，一个紫色非硫光合菌——沼泽红假单胞菌（*Rhodopseudomonas palustris*）和一个海洋单细胞光合菌——聚球藻（*Synechococcus*）菌株 WH8102。此外，该所还为美国能源部感兴趣的环境微生物资源进行调查，并于 2001 年完成了 17 种微生物的测序草图，2002 年，他们又测了另外 6 种微生物基因组。

其他值得一提的微生物测序项目

革兰氏阳性枯草芽孢杆菌是早期完成测序的重要微生物之一。在法国巴黎巴斯德研究所（Institut Pasteur）[32]的领导下，该项目于 1997 年 11 月由欧洲和日本的实验室合作完成。瑞典乌普萨拉大学（University of Uppsala）的一个研究组，1998 年 11 月测序并发表了普氏立克次体（*Rickettsia prowazekii*）的基因组[33]，该病原体能引起流行性斑疹伤寒，营严格体内寄生生活，在进化上被认为是线粒体的起源。铜绿假单胞菌（*Pseudomonas aeruginosa*）PAO1（6.3Mb）是重要的环境细菌，它也是囊肿纤维症（Cystic Fibrosis）和免疫缺陷症患者的病原菌，它的全序列测定由华盛顿大学基因组中心（University of Washington Genome Center）下属的囊肿纤维症基金会（Cystic Fibrosis Foundation）和病原学公司（Pathogenesis Corporation）（均位于华盛顿州西雅图市）合作完成，并于 2000 年 8 月发表[34]。

已完成测序微生物的名录正在迅速增加（图 2）（www.tigr.org）。微生物种类超过地球上所有其他生物，并占全球生物量的极大部分，毫无疑问，微生物基因组的序列数必将继续增长。

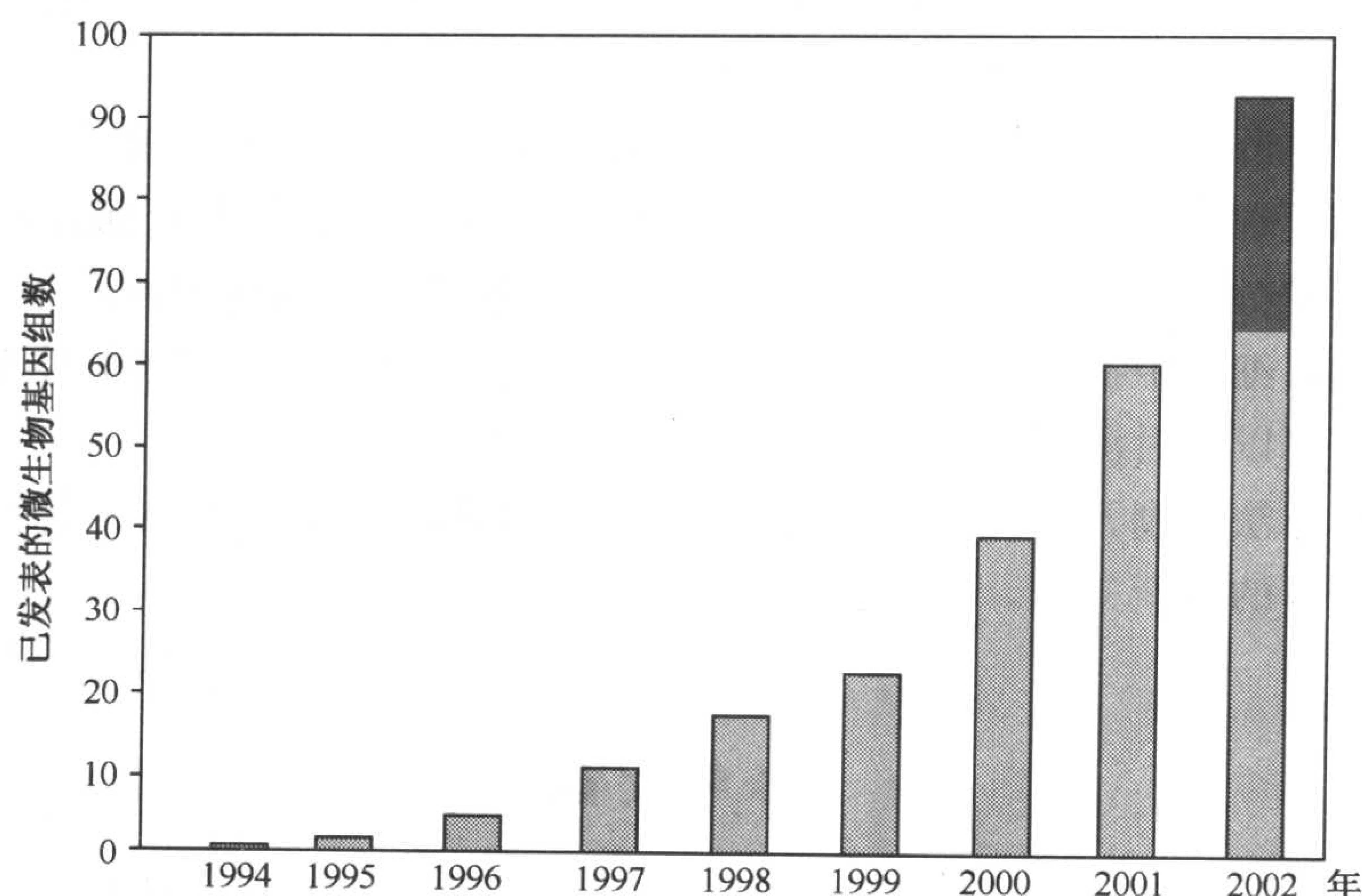


图2 已发表微生物基因组的累计数。数据由 TIGR 微生物基因组数据库提供，截至 2002 年 5 月 30 日。2002 年的数据中包括 28 个已完成测序但还没有发表的基因组，大多数将在 2002 年年底前发表。

测序之后

随着微生物基因组序列的不断积累，科学家们正在积极寻求分析序列的方法。一个生命有机体的基因组不仅包含它的基因序列和调控信息，而且它还是该有机体进化史的文字记录，虽然有时不太好分辨。从序列中提取这种信息不仅要用计算机方法，还要用实验验证。

所有发表的基因组论文都初步尝试了注释基因和记录一些序列特征，比如说重复和复制。在过去的七、八年中，搜寻基因的计算机程序已经发展到了很高的水平。目前 Glimmer（基因定位和内插式马可夫模型，Gene Locator and Interpolated Markov Modeler）[35]，是用于真细菌和古生菌的标准程序。基因功能的预测一般通过与数据库中已知功能的相似基因进行序列对比来完成。尽管如此，每个新基因组序列的面世都会无法与数据库中的条目完全吻合，蛋白数将不断增加，每个新基因组中约有三分之一的基因找不到匹配的数据库条目。

从单基因、基因簇和全基因组等不同水平上比较各个基因组，希望能发现横向或纵向基因转移的例子，以便为它们的进化机制和进化关系提供一些线索。在比较基因组学中，MUMmer[36]计算系统能将两种生物的全基因组排序比对，该系统应用快 3 年了。

寻找调控位点的进度稍慢，例如，目前还没有寻找启动子完全满意的软件，而寻找终止子还比较成功。TransTerm 程序可以在细菌基因组中搜寻依赖 ρ 因子的转录终止子[37]。寻找更加复杂的调控结构的软件还尚待开发。

实验手段主要致力于确定基因的功能以及分析它们在不同生长生理条件下的表达模式。为了使基因失活或“敲除（knock out）”，转座子突变用得最多[38]，它会让我们知道某基因在某特定的生长环境中是否必需，确定基因的功能通常需要更加具体的生化和

结构分析。酵母双杂交系统（two-hybrid system）[39]广泛用于研究基因的产物——蛋白质之间的相互作用，以便提供更多基因功能的线索。

基因表达与调控可以通过全方位的微阵列（microarray）来研究[40]。每个基因的DNA（通常用基因中的某段序列作引物，由聚合酶链反应来合成）都被密集地点在膜上或玻片上组成微阵列，然后分别在标准条件下培养的对照细胞中提取RNA与在特定条件下培养样品细胞中提取的RNA进行标记，标记的RNA可与微阵列杂交，记录杂交量，这样，可以同时测定每个基因的表达水平。

这些不同方法刚刚开始应用，随后几十年的微生物基因组学将为我们提供细胞如何运作的前所未有的崭新知识。

（喻子牛，许朝晖　译）

参考文献

1. Jenkins NA, Kucherlapati RS, McKusick VA. Genomics as it enters the second decade. Genomics 1977; 45:243.
2. Ruddle F. Hundred-year search for the human genome. Annu Rev Genomics Hum Genet 2001; 2:1–8.
3. Luria S, Delbruck M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 1943; 28:491–511.
4. Avery O, Macleod C, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J Exp Med 1944; 79:137–157.
5. Lederberg J, Tatum EL. Novel genotypes in mixed cultures of biochemical mutants of bacteria. Cold Spring Harb Symp Quant Biol 1946; 11:113–114.
6. Hayes W. Recombination in *E. coli* K12: Unidirectional transfer of genetic material. Nature 1952; 169:118–120.
7. Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 1953; 171:737–738.
8. Cairns J. The bacterial chromosome and its manner of replication as seen by autoradiography. J Mol Biol 1963; 6:208, 213.
9. Bachmann BJ. Linkage map of *Escherichia coli* K-12, edition 8. Microbiol Rev 1990; 54:130–197.
10. Sanderson KE, Roth JR. Linkage map of *Salmonella typhimurium*, edition 7. Microbiol Rev 1988; 52:485–532.
11. Piggot PJ, Hoch JA. Revised genetic linkage map of *Bacillus subtilis*. Microbiol Rev 1985; 49:158–179.
12. Maxam AM, Gilbert W. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol 1980; 65:499–560.
13. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977; 74:5463–5467.
14. Sanger F, Coulson AR. The use of thin acrylamide gels for DNA sequencing. FEBS Lett 1978; 87:107–110.
15. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage φX174 DNA. Nature 1977; 265:687–695.
16. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. J Mol Biol 1982; 162:729–773.
17. Smith LM, Sanders JZ, Kaiser RJ, et al. Fluorescence detection in automated DNA sequence analysis. Nature 1986; 321:674–679.
18. Dulbecco R. A turning point in cancer research: sequencing the human genome. Science 1986; 231:1055–1056.

19. Blattner F. Biological frontiers. Science 1983; 222:719–720.
20. Blattner FR, Williams BG, Blechl AE, et al. Charon phages: safer derivatives of bacteriophage lambda for DNA cloning. Science 1977; 196:161–169.
21. Blattner FR, Plunkett G 3rd, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997; 277:1453–1474.
22. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996; 274:546–567.
23. Mewes HW, Albermann K, Bahr M, et al. Overview of the yeast genome. Nature 1997; 387:7–65.
24. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
25. Peterson SN, Hu PC, Bott KF, Hutchison CA 3rd. A survey of the *Mycoplasma genitalium* genome by using random sequencing. J Bacteriol 1993; 175:7918–7930.
26. Balch WE, Magrum LJ, Fox GE, Wolfe RS, Woese CR. An ancient divergence among the bacteria. J Mol Evol 1977; 9:305–311.
27. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 1977; 74:5088–5090.
28. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990; 87:4576–4579.
29. Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 1996; 273:1058–1073.
30. Kotani H, Kaneko T, Matsubayashi T, Sato S, Sugiura M, Tabata SA. Physical map of the genome of a unicellular Cyanobacterium *Synechocystis* sp strain PCC6803. *DNA Res* 1994; 1: 303–307.
31. Kaneko T, Sato S, Kotani H, et al. Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 1996; 3:109–136.
32. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 1997; 390:249–256.
33. Andersson SG, Zomorodipour A, Andersson JO, et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 1998; 396:133–140.
34. Stover CK, Pham XQ, Erwin AL, et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 2000; 406:959–964.
35. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with Glimmer. Nucleic Acids Res 1999; 27:4636–4641.
36. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res 1999; 27:2369–2376.
37. Ermolaeva MD, Khalak HG, White O, Smith HO, Salzberg SLJ. Prediction of transcription terminators in bacterial genomes. J Mol Biol 2000; 301:27–33.
38. Hutchison CA, Peterson SN, Gill SR, et al. Global transposon mutagenesis and a minimal Mycoplasma genome. Science 1999; 286:2165–2169.
39. Bartel PL, Fields S, eds. The Yeast Two-Hybrid System (Advances in Molecular Biology). Oxford, UK: Oxford University Press; 1977.
40. Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. Proc Natl Acad Sci USA 2000; 97:12,170–12,175.

第二部分：作为基因组学工具的生物信息学

2

寻找基因和全基因组比较的工具

Steven L. Salzberg and Arthur L. Delcher

本章将叙述两种基因组分析的计算方法：寻找基因（gene finding）和全基因组比较（whole genome comparison）。通过扫描基因组和分析序列的统计学特征，可以快速、准确地发现原核生物基因组中的基因，通过这种方法可以发现 99％以上的基因。接着通过搜寻附近的调控位点和与其他生物的蛋白质序列比较，进一步优化预测结果，这些步骤可以使用免费软件和数据库自动进行。

全基因组比较是将一种生物与另一种生物的 DNA 序列进行比较，以揭示它们之间所有的相似性和重排、缺失、插入以及多样性等。随着越来越多亲缘关系相近全基因组序列的测定，这种比较已逐渐变成基因组分析的强大工具。这种计算任务可通过利用后缀树（suffix tree），在最短的时间和最小的空间内完成，后缀树是可以在线性时间内建造和搜索的一种数据结构。

引言

从高等真核生物的基因组中寻找基因目前还是个难题，然而，从原核生物（细菌、古生菌、病毒）基因组中寻找基因却相当准确，由于基因组序列计算分析中没有内含子这个巨大障碍。本章叙述的寻找基因软件，可在无人为干预的情况下，发现大多数原核生物中 99％以上的基因。从单细胞真核生物中寻找基因有中度困难，有些单细胞真核生物（如布氏锥虫 *Trypanosoma brucei*）只有少量内含子，因此细菌的寻找基因软件足够发现它们的基因；其他单细胞真核生物（如恶性疟原虫 *Plasmodium falciparum*），由于含有大量内含子，要求使用特殊目的的寻找基因软件，如 GlimmerM[1,2]。与人类和小鼠相比，从这些简单真核生物中寻找基因是简单和准确的，但与细菌相比，还是比较困难的。

现代生物信息学软件对细菌基因组的分析产生了大量准确、丰富的注释信息，为将来研究基因功能打下了基础。在本章我们详细总结了涉及自动寻找基因的计算方法，尤其是 Glimmer 系统[3,4]，还将简述如何使用邻近序列信号，例如核糖体结合位点（RBS），进一步优化这些初步的基因预测结果。

由于近年来完全测序基因组数量的不断增加，包括很多亲缘关系相近的生物，需要能对两个完整基因组进行比较的软件，除了揭示生物体之间的大规模相似性外，这些比较也导致对基因组进化的新发现。最有效的全基因组比较系统，可使用普通台式计算机在一分钟内完成两个细菌基因组的比较。讨论完寻找基因后，本章将介绍支撑这些系统的关键技术 MUMmer[5,6]，并提供用该系统进行序列对比的实例。

GLIMMER：寻找基因的内插式马可夫模型

细菌基因组布满了基因，在一个典型的细菌中，90%DNA序列编码蛋白质序列，它们被较短的间隔序列分开，这些短间隔序列经常含有其他调控序列（古生菌和病毒也有这样的特点，为了讨论方便，我们只谈细菌），此外，细菌编码蛋白质的序列很少被内含子打断。因此，寻找基因的计算方法可采用十分简单的策略就能识别基因，即找出每个比某一固定长度（如500bp）不需改动的可读框（open reading frame，ORF），然后就把从起始密码子（ATG或GTG）到终止密码子间的区域称为一个基因，这称之为简单基因搜寻（simple gene finder，SGF）策略，该策略经常失败，但通过修正，它能变成基因搜寻的一种有效途径。

首先，需要定义ORF，它是在框架内没有终止密码子（TAA，TAG或TGA）的一段核酸三联体序列。我们的兴趣在最长的ORF，所以假设ORF两端各有一个终止密码子，由于ORF被两端终止密码子所界定，所以它们的碱基数目是3的倍数。值得注意的是，不同读框的ORF是重叠的，因为64个DNA三联体密码中有3个是终止密码子，所以预期在非编码区发现终止密码子的频率较高，准确频率数由基因组中GC含量决定。例如，结核分枝杆菌（*Mycobacterium tuberculosis*）的GC含量是66%，它所含终止密码子的频率低于GC含量小于50%的生物。记住这个限制，就可使用SGF算法决定可能最小的ORF长度。Glimmer（后面叙述）通过统计公式计算ORF长度，使某一长度ORF在每一百万个碱基的非编码DNA序列中，随机出现概率只有一次，为简单起见，假定这个长度是600bp。

SGF算法几乎建立了：可快速扫描一个基因组的全部6种可读框，找出全部大于600bp的ORF，称为ORF基因，当然，这会遗漏全部较短基因，将在后面加以说明。首先，SGF需要修改来考虑基因不能重叠的限制，并不完全禁止基因重叠，仅有非常少的基因确实有重叠编码区域，这些重叠通常非常短（一个例子是在TGATG序列中的重叠基因，这里的TGA是终止密码子，ATG是起始密码子）。为保险起见，SGF在初步搜寻中将舍弃全部与其他ORF重叠的ORF。值得注意的是，如果用600bp作为ORF的下限，那么短于600bp的ORF将被忽略，因此，如果一个600bp的ORF覆盖一个短的ORF，那么算法将把长ORF当作一个基因。

SGF有两个严重缺点：第一，它不能识别任何短于预先定义长度（600bp）的基因。已经知道有许多这样的基因存在，当然希望在基因注释中包括它们。第二，当两个或多个ORF重叠时，程序不知如何选择，结果在这个区域内得不到任何基因。对于高GC的生物，这种情况可能在基因组中占很高比例，另外，如果一条DNA单链ORF上缺乏终止密码子，那么就会降低在互补DNA链上存在终止密码子的可能性，因为这容易在互补DNA链上产生重叠ORF。

Glimmer系统试图发现基因组中的全部基因，而不仅仅是那些超过固定长度的基因。Glimmer算法的核心是内插式马可夫模型（interpolated Markov model，IMM），IMM是用于估算某序列是否为编码区域的统计模型。为了建立IMM，Glimmer需要一些训练数据，即一组来自基因组的代表性基因样本，这就像“先有鸡还是先有蛋”的问

题——如何能在使用基因搜寻工具前得到用于训练的基因？用 SGF 可以相当漂亮地解决这个问题。SGF 算法是非常保守的基因搜寻工具，它仅能发现数量有限的基因，却有非常低的假阳性错误率，因此，能用 SGF 得到 GLIMMER 需要的训练数据。GLIMMER 软件包包括称为“long-orfs”的 SGF 算法程序，该程序可以从任何基因组中取得 Glimmer 系统所需的训练数据。对一个典型的 2Mbp 细菌基因组，SGF 可给出 1Mbp 的训练数据。另外，还可用 BLAST 从其他生物的数据库中，搜寻与待分析生物序列有相似性的蛋白质，这些蛋白质也可以作为很好的 Glimmer 训练数据。基因组研究所（The Institute for Genome Research）的自动基因注释系统两种方法都用，先用 long-orfs 运行系统，再用 BLAST 从数据库中搜寻预测蛋白质和收集新的训练数据以后，再次运行系统。

背景知识：马可夫模型

IMM 是基于马可夫链的概率模型技术。首先它有助于解释马可夫链，然后它将显示如何扩展到创建 IMM。一个马可夫链从概念上是一系列状态，每种状态从字母｛A, C, G, T｝中产生一个符号，为了决定哪个符号被产生，状态使用这四个值的概率分布。字母根据概率分布任意选择，因此，如果 $p\ (A)\ =0.2$，那么，此时字母 A 输出概率是 20%，模型进而转移到下一种状态。

一个极其简单的马可夫模型如图 1A 所示，它使用 0.2，0.3，0.4 和 0.1 作为四种核苷酸的概率，该模型可通过简单的运行产生一种 DNA 序列，每次循环产生一个符号。在图 1A 的无意义模型中，序列的每个位置被状态 S_0 所产生，因此，所有位置有同样的概率分布，显然，这不是 DNA 的一种好模型，描述蛋白质编码序列的较好模型应该是三种状态，每种状态描述一个编码位置，这将允许用不同概率描述不同位置，如图 1B 所示。

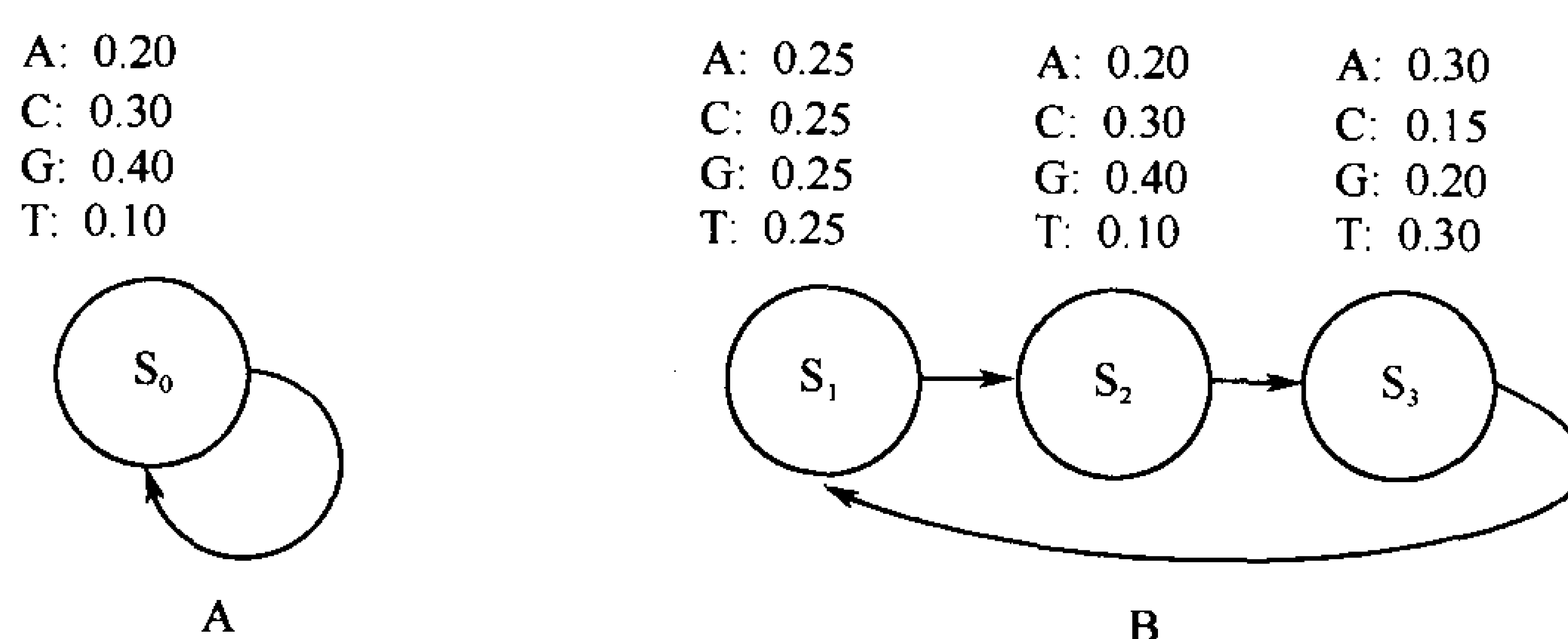


图 1　A. 简单的单状态马可夫链，核苷酸的输出是依靠单个概率分布，而与序列的位置无关。B. 三状态马可夫链，依靠编码位置 1，2，3 的不同概率输出核苷酸；三种状态无限循环。在两种模型中所有的边（箭头所指）为概率 1.0。

显然，我们并不想产生一个人造序列，而是想阅读一个序列并确定它是否是一个基因，为了做到这点，计算模型产生给定序列的概率。像这样反推模型非常简单，操作如下：从序列起始处开始，每种状态“吃进”一个字母并确定一个概率，如果马可夫模型

读到 G，那么，就查询在那个状态时 G 的概率，然后，这个概率与前面所有字母的概率相乘，序列的得分是全部状态所有概率的乘积。实际上，不可能相乘成千上万个小于 1 的数，因为这样的乘积将非常小，例如 10^{-3000}，这就需要特殊数据结构的计算机储存这样的数值，为了避免这个问题，概率能以对数形式储存；对这些对数简单的相加，在数学上相当于概率相乘。

这时，序列只有一个值，它不能说明该序列是否就是一个基因。显然，因为所有概率都小于 1，序列越长，这个值越小。然而，序列是一个基因的概率，并不随序列的增长而变小，相反，它通常会变大，为了确定一个基因，需要将得到的值与其他值比较。为达到这个目的，需要创建一个非基因的模型。

一个可能的模型是任意 DNA，计算整个基因组可以得出 A、C、G 和 T 的概率，然后，就能创建一个无意义的模型（见图 1A）。为了明确下一步做什么，必须弄清楚这些得分的含义，由序列 S 和模型 M 产生的分数是个概率，更准确地说，它是这个模型产生完全相同序列 S 的概率：$p(S|M)$。

已经描述了一个简单随机（或非编码）DNA（如图 1A）的模型，称之为 M_N，编码 DNA 的三状态模型（如图 1B）是 M_C，也知道如何计算 $p(S|M_C), p(S|M_N)$，但是，它们并不是真正所想要的。假设是个好模型（将在下文中修正它们），那么，需要将从编码区模型产生序列的概率 $p(M_C|S)$与从非编码区模型产生序列的概率进行比较。如果计算表明序列更可能是从编码区产生，那么就能确信这个序列是一个基因。

用贝叶斯定律（Bayes'Law）进行计算

$$p(M \mid S) = \frac{p(S \mid M)p(M)}{p(S)}$$

我们将计算给定序列的模型概率，并比较两个模型的值，$p(M_C|S)$对 $p(M_N|S)$，值得注意的是并不知道序列优先概率[$p(S)$]或模型优先概率[$p(M_C)$和 $p(M_N)$]，为解决这个困难，Glimmer 通常假设所有模型都有相同的可能性，或假设一种模型为一个常数因子，而它的概率大于另一种模型，这样，$p(M_C|S)$和 $p(M_N|S)$的总和为 1.0，该值就可决定 $p(S)$的值。

内插式马可夫模型

在前面描述的马可夫模型是零阶马可夫链，一个非常简单的模型，它的任意核苷酸的概率仅由模型的状态决定。Glimmer 使用更加复杂的模型，该模型用多达 8 个碱基计算每个碱基的概率，这 8 个碱基是从有待计算的那个碱基之前的 12 个碱基中选择。

从介绍 Glimmer 的章节中，可以找到对内插式马可夫模型（interpolated Markov model，IMM）的详细说明[3,4]，这里仅作简单的总结。零阶马可夫链的最简单扩展是一阶马可夫链，即每个概率的计算由其他一个位置决定，通常是它前面紧挨的位置。因此，不计算 $p(A)$，而是计算 $p(A|b)$的值，这里，b 是 A，C，G，T 的值，可以把这些概率用于图 1B 中的编码模型。实际经验和数学验证都表明，一阶马可夫链总不比零阶马可夫链差，这可以迅速扩展到用前面两个或三个碱基的二阶或三阶模型计算概率，事实上，数学很容易证明，高阶马可夫模型总是优于低阶马可夫模型。

鉴于这个数学事实，唯一值得注意的是实践。如果概率不准确，那么高阶模型不会优于甚至有可能劣于低阶模型，能快速推算一个典型的细菌基因组，训练数据总数仅足够用于五阶或六阶模型。例如，一个五阶模型要求 $4^6=4096$ 概率，因为全部四种碱基的概率，必须通过五聚体的每种组合估计。如果训练算法（long-orfs，见 Glimmer 部分）产生 1Mbp 的数据，那么，1024 个五聚体中的每个将被平均抽样 1000 次，这将足够估计 4 种概率，每增加一次模型长度，样本数量将减少 4 倍。另外，在细菌基因组中，许多六聚体并没有以足够的比例出现，因此，这些概率的估计不准确。

解决这个问题的方法是：当数据丰富时，使用高阶马可夫模型，反之，使用低阶马可夫模型，IMM 就是依照这一原则设计的。在 Glimmer 系统内部构建了 9 种马可夫模型，从零阶到八阶，然后，用训练数据进行计算统计。在这里 Glimmer 针对数据的丰富程度，对每一阶马可夫模型进行计算，最大到 8 阶。本质上这些统计数据告诉 IMM，是否有足够数据准确估计某一给定多聚体的 4 种概率（A，C，G，T）。然后，系统选择可能最长的多聚体给序列中的每个碱基赋值，这些值是在一个 ORF 中所有概率的乘积。

打分

对含有 ORF 基因组任一段区域，细菌基因搜寻工具要做的主要决定是这个区域是否含有基因，如果是，哪种可读框含有这个基因。有 7 种选择：每条 DNA 单链的三种可读框中可能有一个基因，或者都没有基因，默认这些选择中假设只有一个正确，虽然可能存在重叠基因，但这种情况很少，一般仅有少数几个核苷酸重叠，可以设置 Glimmer 去发现重叠基因，默认设置为允许少量重叠。

Glimmer 构建几个 IMM，把它们都给 ORF 打分，并决定哪些是真基因，先构建第一、二、三编码位置的 IMM，在给 ORF 打分时，每个位置用不同的 IMM，这称为“三周期（three-periodic）”马可夫模型，因为系统是在 ORF 中每三个位置的三种模型间循环。在给一个特定 ORF 打分时，系统把在那个 ORF 可读框的三周期 IMM 中得出的概率相乘。在给基因组中的一个区域打分时，系统构建 6 种不同可读框（当然有些包括终止密码子），并给它们打分。对于第 7 种选择（无基因），系统用类似图 1A 中的随机 DNA 简单模型。所有这些分值加以计算，如果一个模型的分数高于全部模型 90%（这个阀值可被用户调整），那么这个模型将会“胜出”。因此，Glimmer 只有在一个可读框内（in-frame）的 ORF，胜过其他竞争模型后才被确定为一个基因。

Glimmer2.0 系统增加一些步骤，用于解决基因内重叠的问题，最终输出结果是显示全部 IMM 分数的详细基因表格，包括那些不被系统认为是基因的 ORF 的分数，还有显示全部预测基因和它们在基因组中位置的一个简表。Glimmer（1.0 和 2.0）的很好优点是把 IMM 模型作为独立的模块单独输出，它允许用户以一个生物的数据训练 IMM，然后，将其应用到另外的生物中去搜寻基因。这条途径非常有用，因为很多时候科学家想从较短的 DNA 片段中发现基因，但又没有合适的训练数据，在这种情况下，就可用亲缘关系相近的生物进行训练，然后，再在序列片段上运行基因搜寻模块。

研究结果[4]显示，在没有人为干预的情况下，Glimmer2.0 可以发现大多数细菌基

因组99%以上的基因，当然，也存在少量假阳性错误率，但这个比率很难准确估计。大多数基因组中约有5%～15%的预测基因，不能与已发表基因注释相吻合，但是，没有进一步证据证明这些基因中有多少是真正的未知基因，有多少是错误的预测。至今，在大肠杆菌这种最广泛研究的细菌基因组中，还有新蛋白质编码基因被发现，因此，对于其他基因组，要想发现它们的全部蛋白质还需要更多的研究，例如，Wassarman及其同事[7]从大肠杆菌已被注释为基因间隔区的一个序列中，发现了6个新蛋白质序列，而所有这些蛋白编码序列都曾被Glimmer准确地预测为基因。

用核糖体结合位点识别起始密码

改进细菌基因搜寻的另一个工具是用核糖体结合位点（ribosome-binding site, RBS)，RBS是一段位于大多数基因起始密码子上游的短序列，称为Shine-Delgarno序列（SD序列)，它与16S核糖体RNA的3′端互补，并通常以AGGAG（在大肠杆菌和许多其他细菌中）形式出现。我们已经利用RBS开发了一种算法，改进Glimmer、GeneMark[8]和其他基因搜寻工具对起始位点的预测。RBSfinder程序可从Glimmer主页（www.tigr.org/software/glimmer）上免费得到，并作为Glimmer程序的后加工程序来运行[9]。用户只要简单地输入基因组及其Glimmer的预测结果，RSBfinder就以SD序列作为指导，试图找出每个基因的RBS。

如果一个给定基因组16S核糖体RNA是未知的，那么，程序可用吉布斯抽样（Gibbs sampling）的一种变化方式解决这个问题。其基本原理如下：首先，假设Glimmer输出的许多基因都有正确起始位点，RBS就应该出现在预测起始位点上游10～15个碱基的范围内。以此为指导，算法便从所有预测基因的上游抽出一小段序列（窗口序列)，并反复搜寻这些序列以得到最普遍的序列模体（motif)，用户可以调整窗口序列和模体序列的长度，抽样算法不断修改这个模体序列，直到得到一个稳定的模体（通常不需要很长时间)。最后，使用用户提供的RBS或吉布斯抽样方法得到的模体，算法扫描全部预测基因，并定义一个描述RBS中每个碱基概率的位置权重矩阵（position weight matrix)[10]，这个矩阵在功能上是零阶马可夫模型，它可以给任何假定序列评分。

使用位置权重矩阵，算法扫描各个基因并尝试修改任何没有较高分值RBS的起始位点，它会搜寻每个起始密码子的上游和下游，以发现有更好RBS位点的起始密码子。如果在从另一个起始密码子的适当距离发现了一个好位点，那么，它将在保持原有可读框的前提下，把起始位点改到新位置。然后，程序给出一个完全的基因列表，包括基因旧起始密码子和调整后的起始密码子，以及预测RBS的位置和序列。

可以把Glimmer所有预测蛋白质与蛋白质序列数据库进行比较，进一步确认预测结果，并在必要时加以修正，包括这个步骤在内的全部注释过程，将在第3章中讨论。

全基因组比较

基因组研究所1999年完成了结核分枝杆菌CDC1551的基因组后，马上遇到的问题

之一是与亲缘关系相近的实验室菌株 H37Rv[11]的比较。该菌株的基因组一年前才发表，其基因组长接近 4.4Mbp，当时，还没有软件可用于分析如此长的序列。这两个基因组平均 99%相同，仅有约 1000 个单核苷酸和少量插入序列不同，当时在计算上所遇到的问题，是将两个序列排列对比从而发现它们所有不同之处。

比较两个基因或一个基因与大规模基因数据库的问题已经解决了一段时间，它的有效应用工具已出现在最新 BLAST［基本局部联配搜寻工具（basic local alignment search tool，BLAST）］[12]和 FASTA[13]系统中。在第二个结核分枝杆菌的基因组鉴定前，没有人需要一个程序去比较以兆碱基计的序列。BLAST 和 FASFA 的算法要求计算时间的平方（即所需要的时间是所输入序列长度平方的函数），它仅仅适合如同蛋白质序列的较短输入，但对全基因组则相当昂贵。因此，我们研究组创建了基于最大特异性吻合（maximal unique match，MUM）和后缀树的一种新算法，称为 MUMmer。

如果两个基因组非常相近，那么，它们的许多 DNA 序列应该非常吻合，而这些序列在两者中有相同的次序和方向。MUMmer 就是为了发现全部这样的序列并把它们聚积成较大的簇（cluster）而设计的，吻合的序列是 mums，它定义为精确吻合两个基因组的子序列，它在每个基因组中只存在一个，并具有最大可能的长度，这个最后要求意味着，如果在序列的任一端扩展一个碱基，那么，它们就不再彼此吻合了，如图 2 所示。

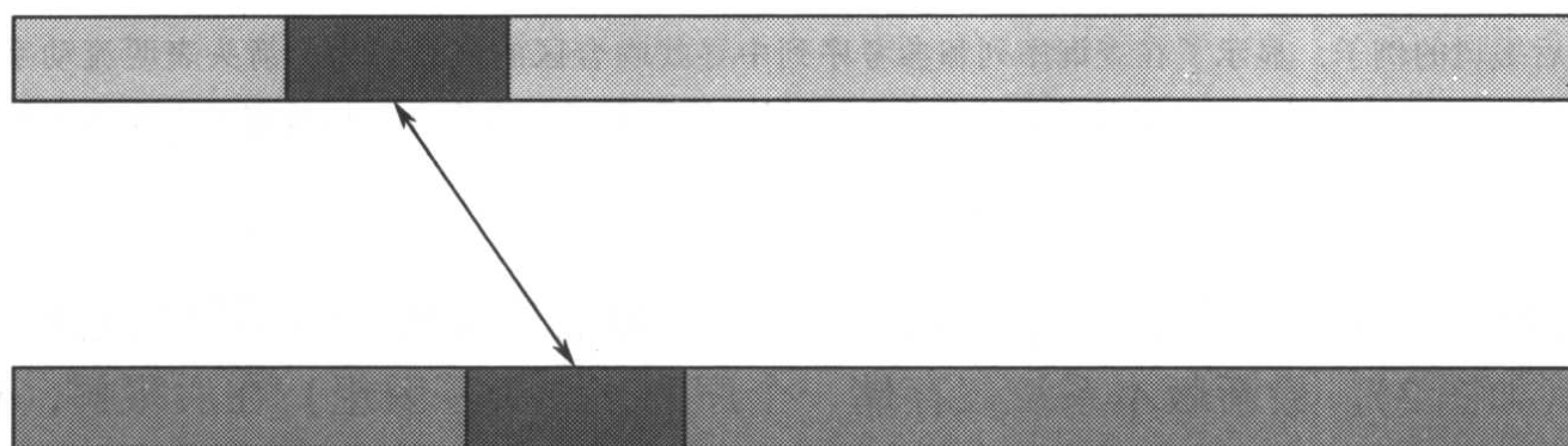


图 2　两个基因组由上、下两个方框表示，最大特异性吻合（MUM）在图中表示为深灰色。mums 序列在每个基因组中只出现一次，它不能在没有错配的情况下向任一端扩展。

MUMmer 算法的精深步骤涉及到毫无遗漏地发现两个基因组的全部 mums，然后，它们分成具有相似性较长区域的簇。这些集簇步骤可被用户控制，用户能指定在簇内 mums 之间的最大距离，MUMmer 还包括另外两个软件包，它们可用于比较部分组装的基因组，也可用于基于蛋白质序列相似性的比较（解释如下）。首先，我们概述后缀树（suffix tree）的建立和搜寻吻合序列的算法，这仅仅是个概述，全部细节请看原始文献[5]。

后缀树是含有一个特定序列全部子序列的数据结构，对序列 atgtgtgtc $ 后缀树的例子如图 3（$ 是用来方便地标明序列末端的特殊字符），树中的节点代表序列中的位置，每片叶子节点（最低端节点）代表一个后缀，它定义为起始于任意位置并一直延伸到序列末端的一段序列。在图中，叶子（方框）标明参考序列中后缀开始的位置，边线用原输入序列的子序列标明，同一节点下的边线，必须用起始不同位点的字母标明，每个不是叶子的节点下，必须至少有两条边线。为了重建任何叶子节点的后缀，只需简单地从

根到叶向下追寻它的路径，连接所有遇到边线上的标注。虽然长度为 n 的序列含有 $(n^2+n)/2$个子序列，但是后缀树能够把所有相同子序列组织到一个数据结构中，而且该数据结构最多只有 n 个叶子和 $n-1$ 个非叶子的节点。

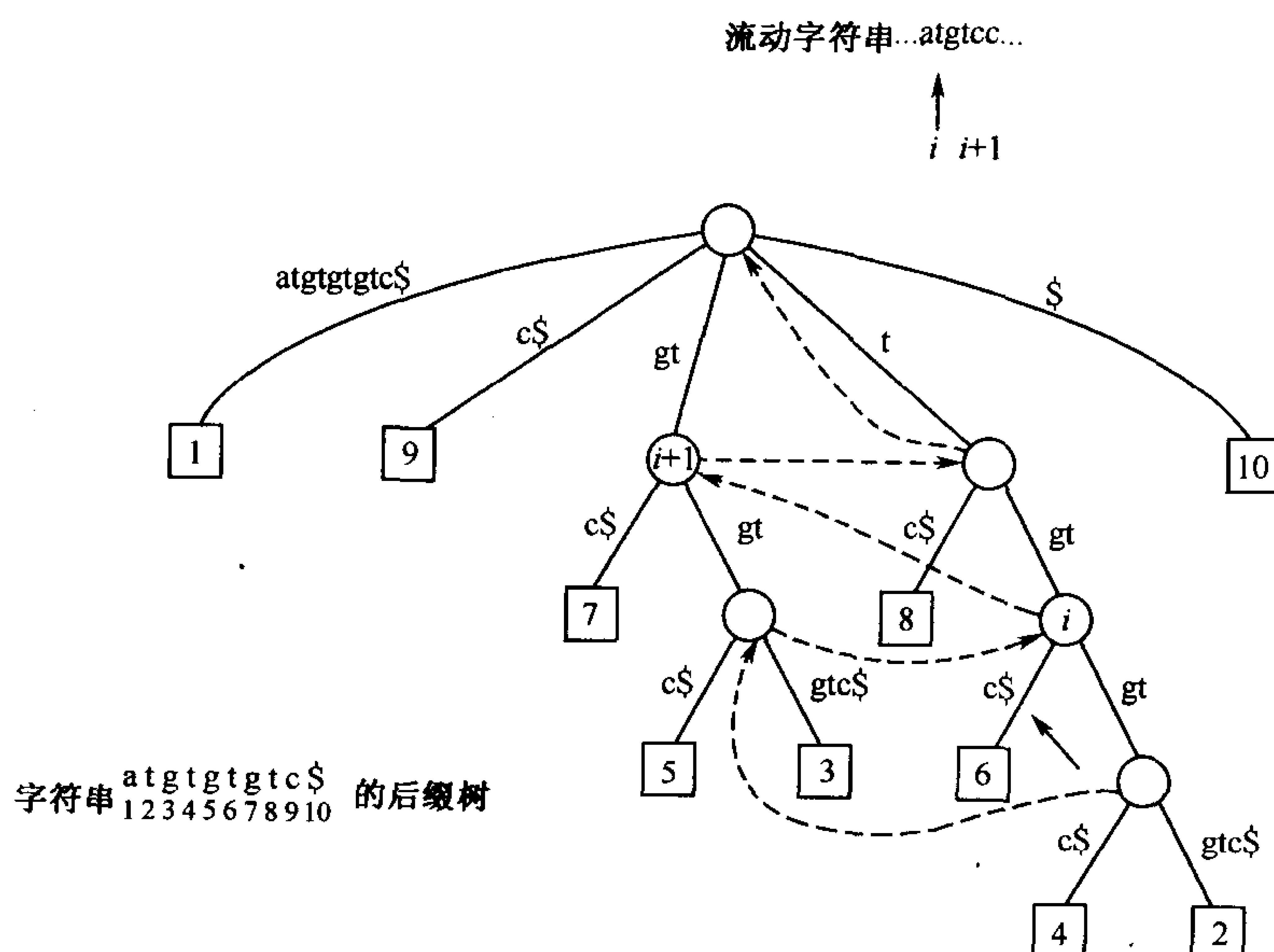


图 3　一个后缀树的例子，显示了在查询序列和参考序列中寻找吻合区的流线行为。箭头表明流动字符串（查询序列）的某一位置与在树中另一个箭头所指位置相吻合。在箭头下方的节点 6 相对应于在参考字符串中的位置 6，表示在查询序列和参考序列中都含有序列 tgtc。

最初 MUMmer 算法[5]的创建包含两个输入序列的后缀树，然后寻找它们之间的全部 MUM（如图 3)。最新版本系统仅存储一个序列（一个基因组）在后缀树，称为参考序列。第二个序列“流过”后缀树，意味着让这个序列通过这个树，标记全部能够吻合的位置，这个方法是由 Chang and Lawler[14]引进，完整的描述见 Gusfield 的著作[15]。因为，这个后缀树含有第一个基因组的全部序列，所以通过“流动法”可以发现第二个基因组中全部吻合序列。这与最初算法的主要不同，是特异性吻合仅仅针对第一个基因组，而不是针对两个基因组。如果想得到针对两者的特异性吻合，最初算法也包括在这个软件包中。

这个流动算法的结果是，在与查询序列长度成比例的范围内，查询序列与参考序列中的最大吻合都可以找出来。这种算法的好处是，一旦建好参考序列后缀树，任意长度的多个查询序列都能流过这个树。实际上，使用这些程序比较了两套人类基因组的全序列（每个接近 27 亿个碱基），我们用其中一套序列中的每个染色体作参考序列，并让另一套全部基因组流过它（A. Halpern，个人通信，2002)。流动算法也能大大加快比较小基因组或比较单独染色体与大的多染色体基因组的速度，只要把较小基因组作为参考序列，然后把较大基因组进行查询，所需数据空间的大小将与较小序列成比例。

MUMmer 序列对比的主要输出结果，是所有精确吻合序列的一个表。最简单观察序列对比的方法，是绘一个简单的点阵图，用图中的对角线指示吻合区域。图 4 显示了两个幽门螺旋杆菌（*Helicobacter pylori*）菌株的序列对比（在一个标准的运行 Linux 的

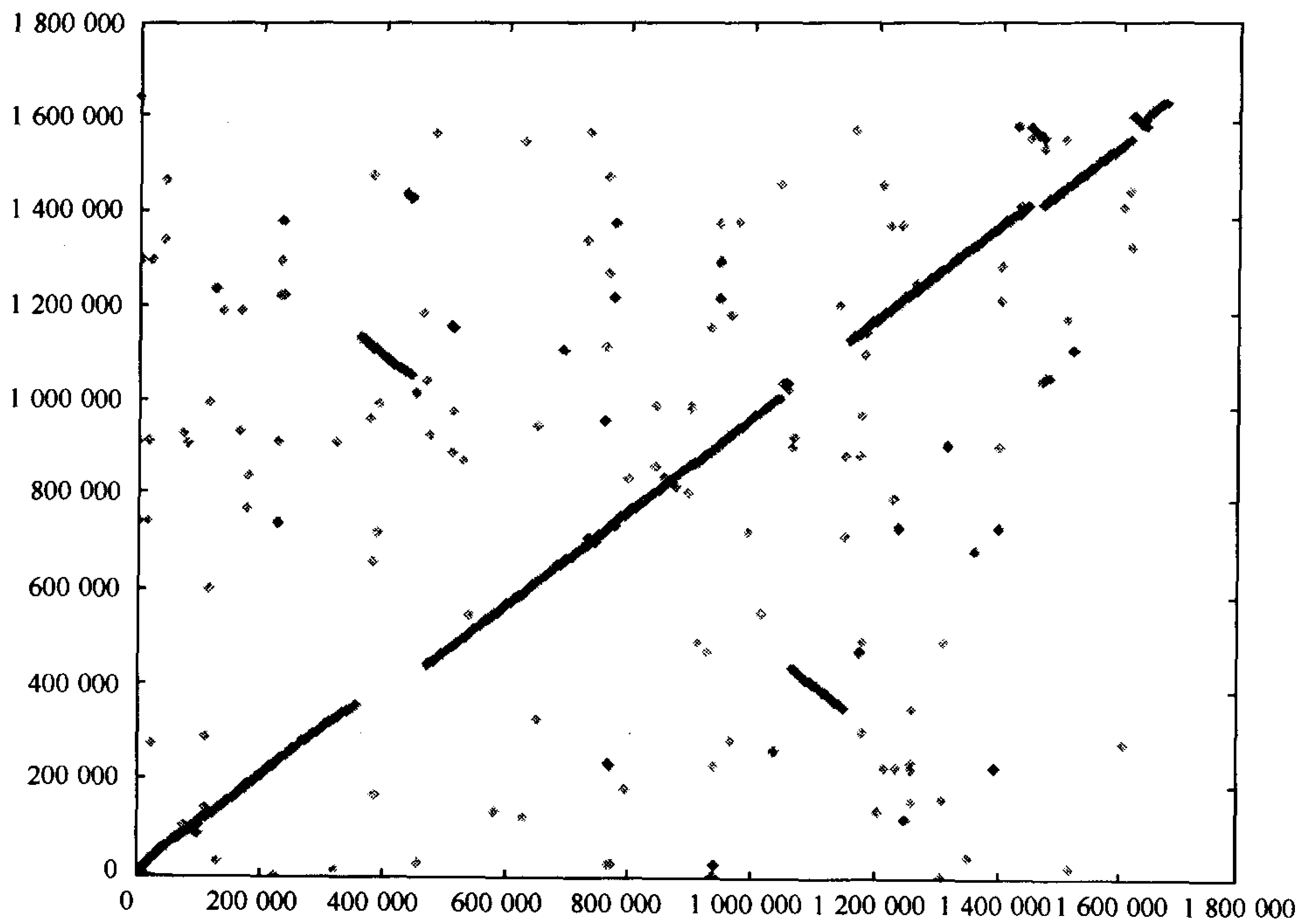


图 4 两个幽门螺旋杆菌对比的全部 MUM 点阵图。横坐标，幽门螺旋杆菌菌株 26695；纵坐标，菌株 J99。长对角线代表可以对齐区域，斜率为 - 1 的对角线表示大规模染色体倒位。

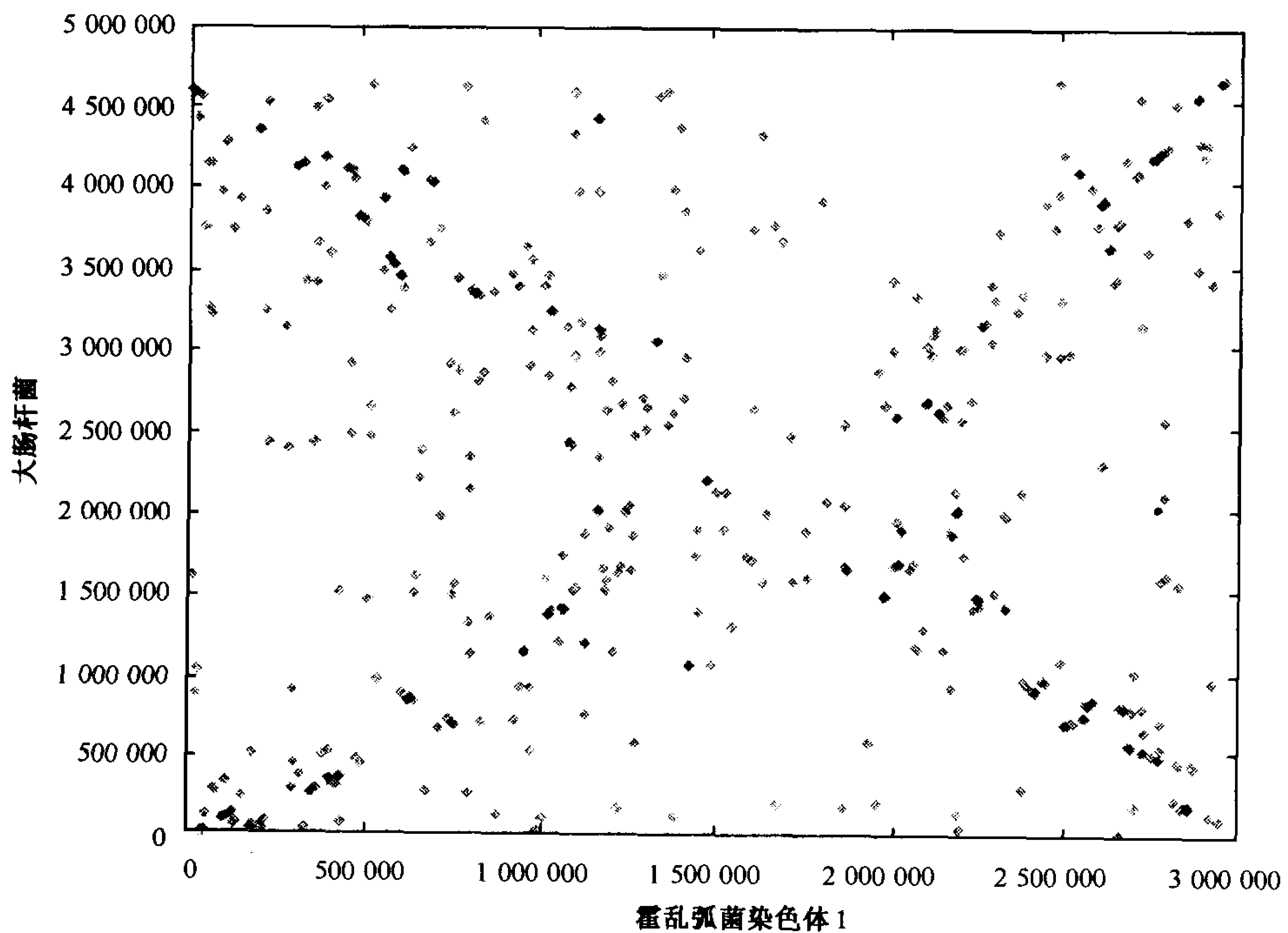


图 5 大肠杆菌和霍乱弧菌间大于 20bp 的所有 MUM 点阵图；展示了由大规模染色体倒位所形成的 X 形排列；所有倒位均围绕复制原点而对称。

台式计算机上，可以在 20 秒内给出该结果），图中十分清楚地显示了这两个菌株彼此间大量的相同序列，并且也突出了 4 个倒置的染色体大片段。复制原点位于第三和第四个倒置区域之间[16]，因此，这些倒置片段是围绕复制原点对称分布。正如较大规模研究所显示的那样，细菌染色体经常围绕复制原点进行倒位，对许多菌株的比较都显示出“X”排列趋势。幽门螺旋杆菌的菌株刚刚开始形成 X 样式，当进行许多倒位后，保守片段（在点线图中的对角线）会变成越来越多的小段，X 样式就会越来越明显。大肠杆菌（*E. coli*）和霍乱弧菌（*Vibrio cholera*）的序列对比例子如图 5 所示，它们有较远的亲缘关系，但 X 样式仍依稀可辨。

最新版本的 MUMmer 可以从我们的网址（www. tigr. org/software/mummer）中获得，它包括一个帮助用户导航输出结果的阅读器（DisplayMUMs）。

部分鸟枪法数据组合的比较

因为后缀树算法在时间和空间上效率都高，它只需比细菌基因组多一点要求，就能用于大的真核生物染色体的比较。它已经用于模式植物拟南芥（*Arabidopsis thaliana*）全部染色体的比较，并导致了对其全基因组的近期复制事件的重大发现[18]。它甚至用于更大规模人类基因组染色体彼此间的比较，发现了数量众多涉及大量染色体的大规模古老复制事件[19]。

除了用于比较完整的基因组和染色体外，MUMmer 还特别适用于比较不完整的基因组。随着基因组完成数量的增加，未完成基因组的数量也在增加，其中有些基因组全序列需要很多年才能完成（如果真有那么一天的话），一种原因是关闭全部间隙序列，并完成一个基因组的测序比鸟枪法测序更加困难和耗时，而鸟枪法能以快速和高通量形式进行。许多基因组已完成了一定覆盖率（1× 到 8×）测序，而且这些信息已经公布，但没有人愿意进一步完成它们。

MUMmer 能像比较全基因组那样，轻松地比较不完全基因组。输入系统可以是代表不同完成阶段基因组的多个 FASTA 文件中的一对，它们中可以有一个或两个未完成序列。系统可以像处理全基因组那样，快速地比较彼此的所有 DNA 重叠群（contiguous segment of DNA，contig；由基因组组装程序产生的一些 DNA 连续片段）。这个系统包括一个独立软件，称为 NUCmer，它能将所吻合的部分聚集成簇，并鉴别对应 DNA 重叠群的次序和方位。例如，如果基因组 A 的三个小 DNA 重叠群与在基因组 B 中的一个大 DNA 重叠群相对应，那么，NUCmer 将显示它们的吻合位置和方位。

对于亲缘关系比较远的物种，由于序列进化产生的分歧，MUMmer 对比也许不能检测这些 DNA 序列的相似性。但是，因为蛋白质序列可以在相当长时间内保持相似性，所以我们另外设计了一个程序 PROmer，它是在基于蛋白质相似性的基础上，用 MUMmer 来对比基因组（或部分基因组序列）。PROmer 用 6 种可读框翻译所有输入序列，产生全部大于某一预先设定的最小长度（用户可以方便地设置这一长度）的蛋白质序列。然后，系统将连接这些序列，寻找所有 mums（用氨基酸字母和“短语”），再将这些吻合部分还原到相应的 DNA 序列中，产生的速图谱使用户可快速发现任何吻合重叠群的顺序和方位。

我们已在最近的工作中使用它，来对比恶性疟原虫（*Plasmodium falciparum*）和约氏疟原虫（*Plasmodium yoelii*），它们分别引起人和老鼠的疟疾。恶性疟原虫的工作当时近乎完成，而约氏疟原虫的工作由于经费紧张仅仅完成了五倍覆盖率，产生了数以千计的重叠群。PROmer 系统能把这些大量重叠群，定位到恶性疟原虫的基因组上，并创建了基于共线性图谱的“假重叠群（pseudo-contig)”。这些假重叠群是除了顺序和方位以外，没有其他任何信息的一些孤立片段，根据它们的假定顺序，返回去运行聚合酶链反应（PCR)，把它们中的许多片段连接在一起，证实了起初用序列对比为基础而创建图谱的可靠性。在这个项目中，PROmer 和 MUMmer 的速度非常重要，因为两个物种的基因组序列都要重新组合很多次，每次组合都要求全部对比重新运行。

致谢

这些研究工作部分由美国自然科学基金会基金（主持人 S.L.S, 项目编号 IIS-9902923；主持人 A.L.D, 项目编号 IIS-9820497）以及美国国立卫生研究院基金（主持人 S.L.S, 项目编号 R01-LM06845）资助。

（喻晓辉，刘　超　译）

参考文献

1. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H. Interpolated Markov models for eukaryotic gene finding. Genomics 1999; 59:24–31.
2. Pertea M, Salzberg SL. Computational gene finding in plants. Plant Mol Biol 2002; 48:39–48.
3. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Res 1998; 26:544–548.
4. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999; 27:4636–4641.
5. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res 1999; 27:2369–2376.
6. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 2002; 30:2478–2483.
7. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev 2001; 15:1637–1651.
8. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 1998; 26:1107–1115.
9. Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics 2001; 17:1123–1130.
10. Claverie J-M, Audic S. The statistical significance of nucleotide position-weight matrices. Comput Appl Biosci 1996; 12:431–440.
11. Cole ST, Barrell BG. Analysis of the genome of *Mycobacterium* H37Rv. Novartis Found Symp 1998; 217:160–172, discussion 172–167.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol 1990; 215:403–410.

13. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 2000; 132:185–219.
14. Chang WI, Lawler EL. Sublinear expected time approximate string matching and biological applications. Algorithmica 1994; 12:327–344.
15. Gusfield D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. New York: Cambridge University Press; 1997.
16. Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF. Skewed oligomers and origins of replication. Gene 1998; 217:57–67.
17. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 2000; 1: research11.01–09.
18. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 2000; 408:796–815.
19. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291: 1304–1351.

3

TIGR 的细菌基因组注释

Owen White

引言

基因组注释最主要的目的是尽可能准确地指出基因组中所有基因的功能。在对许多微生物基因组进行注释的过程中，基因组研究所（The Institute of Genome Research, TIGR）已经开发并利用了很多自动注释方法和工具来分析全基因组序列。除了利用这些自动的方法和工具外，一群训练有素、具有博士学位的科学家们还对序列进行了人工整理（manual curation）。在过去 9 年里，用于基因组注释的工具不断地演化，把很多功能结合在一起，其中包括自动识别基因、识别非编码特征序列（如调控位点和重复序列）、发现基因在数据库中的吻合条目（database match），以及指定基因的作用类型（role category）。下文是对原核生物常用注释方法的纲要性总结，介绍之前，简要地谈谈另一个重要问题：细菌注释的一致性。

细菌数据的一致性

使用细菌基因组项目数据的科学家们都知道，利用这些数据有很多问题，用计算方法给某个基因指定功能不如用实验方法可靠，我们是否应该责备计算机指定法（*in silico* assignment）呢？也许是吧，因为除非基因的生化功能和细胞活性得以证实，用计算方法所指定的基因功能只能是个假说，假说有时是错误的。尽管用实验指定基因功能的方法比较可靠，但还是不能解决所有问题，因为，注释使用的语言和描述方法各不相同，注释时，可以用多种概念描述数据类型，像俗名 [common name，（如酒精脱氢酶)]、遗传名 [genetic name，（如 *adh*1)] 或酶学委员会（Enzyme Commission, EC）的数字命名法（如 1.1.1.1）。在描述基因功能时，也可以用代谢途径的成员、翻译起始点以及转运蛋白的类型等归类。

用这些不同方法描述数据类型就带来了一个问题，例如，描述基因在细胞内的生物功能时，TIGR 采用起源于参考文献 [1] 的分类系统；另外还有其他功能分类系统，例如，由 Ashburner 及其同事创建的“基因本体论（Gene Ontology)”系统[2]。不同实验室采用不同系统描述生物功能的数据类型，这给基因组比较设置了障碍。

另一个问题是在指定数据类型时采用的标准各不相同，例如，给基因指定功能是注释的中心任务，几乎所有发表原核生物的基因组都少不了这方法的内容。遗憾的是，这种指定工作没有一个共同标准。图 1 所举的例子中，几项不同 GenBank 记录（record）采用了许多同义词给一个基因指定俗名，在这些 GenBank 的记录中，“/gene”的注释语各不相同；这似乎微不足道，但是，全面检索具有相同功能基因的尝试却因此而无法实

现。更糟糕的是，这些注释不一致的基因，又被用来指定新基因组中的基因，使错误不断地转移和放大。

A

Escherichia coli
/gene="purU"
/EC_number="3.5.1.10"
/function="enzyme; Purine ribonucleotide biosynthesis"
/product="formyltetrahydrofolate deformylase; for purT-dependent FGAR synthesis"

Campylobacter jejuni
/gene="purU"
/EC_number="3.5.1.10"
/product="formyltetrahydrofolate deformylase"

Methylobacterium sp. CM4
/gene="PurU"
/product="purU protein"

Synechocystis sp. PCC6803
/gene="purU"
/product="phosphoribosylglycinamide formyltransferase"

Rhodospirillum rubrum
/gene="purU"
/product="formyltetrahydrofolate deformylase"

Halobacterium sp. NRC-1
/gene="purU"
/product="formyltetrahydrofolate deformylase"

Helicobacter pylori, strain J99
/gene="purU"
/product="FORMYLTETRAHYDROFOLATE HYDROLASE"

Helicobacter pylori
/gene="HP1434"
/product="formyltetrahydrofolate hydrolase (purU)"

Streptomyces coelicolor
/gene="SCD10.35"
/product="putative formyltetrahydrofolate deformylase (fragment)"

Mycobacterium tuberculosis H37Rv
/gene="purU"
/product="purU"

Corynebacterium sp.
/gene="purU"
/product="10-formyltetrahydrofolate hydrolase"

图 1A　GenBank 中原核生物注释的基因类型不一致。A 图中的文字代表在 GenBank 中，描述某一基因通常所列的一部分信息。这里用一些"标签（tag)"（如/gene）作为计算机的可解析区（parsable field）来存储注释信息。这里所列信息显示在不同微生物中所用标签的相异性。有几处用"/gene"标签存储遗传名 purU，但有的微生物不是。"/product"区一般用来存储俗名（如"formyltertrahydrofolate deformylase"），但有几处也用来存储遗传名。EC 数字名虽然在大肠杆菌中明确标注，但它却经常出现在 GenBank 记录的其他部分，如在 B 图中，它被标注在基因俗名区。对某一特定数据类型的注释也存在不一致现象，如"/product"标签中的不同俗名。

必须认识到，使用不同注释信息不见得不对，可以用图书馆的书比喻注释信息，大

B

32 genes for DNA polymerase III, alpha subunit (dnaE).

3 DNA polymerase III
1 DNA-directed DNA polymerase (EC 2.7.7.7) III alpha chain
2 probable DNA-directed DNA polymerase (EC 2.7.7.7) III alpha chain 3
1 DNA-directed DNA polymerase (EC 2.7.7.7) III alpha chain, spliced form 1
1 probable DNA polymerase III, alpha chain
18 DNA polymerase III, alpha chain
1 putative DNA polymerase III alpha chain
3 DNA polymerase III alpha subunit
1 DNA polymerase III, alpha subunit
1 DNA pol III alpha

EC number: 18
genetic names: 6
functional info: 13

---

18 genes for homoserine acetyltransferase

2 probable homoserine O-acetyltransferase
2 putative homoserine O-acetyltransferase
9 homoserine O-acetyltransferase
2 homoserine-o-acetyltransferase
1 probable metA protein
2 homoserine O-trans-acetylase (yeast)

EC number: 3
genetic names: 6 (metA, met2)
functional information: 2

---

20 genes for glycogen operon protein (glgX).

5 glycogen operon protein GlgX
1 glycogen operon protein GlgX (EC 3.2.1.-)
1 longer ORF due to differences from ECOGLG
4 glycogen debranching enzyme
2 putative glycogen debranching enzyme
1 probable glycosyl hydrolase
1 putative glycosyl hydrolase
2 glycogen operon protein (EC 3.2.1.-) glgX
1 probable glycogen hydrolase (debranching)
2 glycosyl hydrolase family protein

genetic names: 8 (treX, glgX, glgX2)
EC number: 6
functional information: 3

图 1B 图中列举了从 GenBank 中用关键词在 Entrez 文献检索服务器中搜索到的一些关于 DNA 聚合酶 III 的 α 亚基 *dna*E（DNA polymerase III, alpha subunit, *dnaE*）、高丝氨酸乙酰转移酶 *metA* 或 *met2*（homoserine acetyltransferase, *metA* 或 *met2*）以及糖原操纵子蛋白 *glgX*（glycogen operon protein, *glgX*）的俗名。表中显示每个名字在 GenBank 注释中出现过的次数。有 EC 号、遗传名以及列举功能信息（如这些基因的生物作用）条目出现过的次数也列出。

多数图书馆都按照一定操作程序把书安排在书架上，这些放书程序使得每个图书馆都能在需要时很快把书找着。如果把两个图书馆的分类目录下的书进行比较，常常会在同一目录下看到不同的书，即使这两个图书馆的书都一样。可能一个图书馆采用杜威十进制

分类法（Dewey Decimal），而另一个采用美国国会图书馆（The Library of Congress）分类法，这就是数据类型相异的例子。此外，即使两个图书馆都采用杜威十进制分类法，同一本书也可能指定为不同杜威十进制数，这就是指定标准相异的例子。每个图书馆在书架上摆书的位置并不一定不对，只是存储书没有一致性，削弱了图书馆间相互检索的意义。

同样的道理，这里并不是说原核生物全基因组注释有错，只有在使用完全相同数据类型和实施完全相同指定标准的情况下，才有可能对各种原核生物的基因组进行成功比较，当前在公共领域的全基因组注释工作中还缺乏这样的一致性。

为细菌蛋白质指定功能

为了减少在指定数据类型时产生太多的相异性，TIGR 制定了一套给微生物基因组注释时使用的标准操作程序（standard operational procedure，SOP），如表 1 所示，这套 SOP 可以指定很多数据类型，如编码区、功能、俗名、生物作用、创建蛋白质家族、分析疏水性、鉴定可能的转运蛋白、重复序列以及结构核糖核酸（RNA）等。

注释分两步进行，先用自动注释给基因初步指定功能，然后再进行人工整理。在自动注释时，先对由公共领域中存档的所有蛋白质组成的非冗余（nonredundant）数据库进行搜索，搜索工具是 Basic Local Alignment Search Tool（BLAST）-Extend-Repraze（BER），该程序先对每个蛋白都在非冗余数据库中做一次 BLAST 搜索[3]，并把所有十分吻合的结果存储下来。然后，再用修改的 Smith-Waterman 方法[4]，把每个蛋白再与它的 BLAST 吻合条目进行序列对比。为了分辨可能的移码突变（frameshift mutation）和点突变（point mutation），序列都沿预测基因编码区的上游和下游各延伸了 300 个核苷酸。

自动注释程序会找出 BER 搜索的最佳吻合条目，并以它为准给基因指定一个俗名、基因符号和 EC 号，执行该任务的软件以高一致性（high-percentage identity）（至少 35%）为标准，对基因序列进行全长对比（目标序列的 80%以上），如果出现一个以上的吻合条目，程序就会选择符合 TIGR 命名习惯的条目为此基因定名。

该程序也对 TIGRFAM 数据库进行检索[5]（下文亦有介绍），假如 BER 选择结果是另一微生物的假定蛋白（hypothetical protein），或者根本找不到吻合条目，程序就会返回到 TIGRFAM 条目中寻找具有微羽同源性的序列。如果成功，程序就根据 TIGR-FAM 的规则给目标蛋白一个蛋白家族的名字，例如“转录调控因子，TetR 家族”，如果目标蛋白与另一微生物的假定蛋白吻合，却在 TIGRFAM 中找不到吻合条目，那么它就命名为“保守假定蛋白”；既没有与 TIGRFAM 吻合的结果，又没有与 BER 吻合结果的蛋白质就称为“假定蛋白”。

表1 TIGR基因组注释的现行标准操作程序

整理注释（curated annotation）。见正文。

基因搜寻。暂未定名的DNA序列，先由Glimmer[8]进行搜索，Glimmer可以指出某段序列成为潜在编码区的可能性，Glimmer识别已知基因的灵敏度约为99%（见第2章）。

HMM搜索。预测出编码区再进行TIGRFAM HMM和Pfam数据库搜索，图形化的用户界面可以快速分析并指定暂未定名基因的功能。

疏水性、跨膜区域和信号肽。通过分析数据库中的吻合序列识别信号肽和跨膜区域。

基因间隔区分析。Glimmer可以预测编码区，然后再搜索它们所编码的蛋白质与已知蛋白质的相似性。在有些情况下（如种间转移的基因组区域），基因的组成很特殊，以致无法被Glimmer识别。为了纠正这种错误，不含吻合序列的可读框或根本不含可读框的区域被再次扫描。这些“基因间隔区”的所有六个可读框都被扫描，希望能发现吻合序列，如果发现，就通过对该区域进行双序列比对确定可读框终止点。注释员对这些候选基因进行分析后，再给它们作最终注释。

插入序列元件。插入序列（insertion sequence, IS）元件较小，结构简单，只含一个或几个ORF，其中有一个或两个ORF编码转座酶。注释员对IS元件左右两端的序列进行分析，人工识别它们的边界。IS元件左右末端序列是转座酶的作用位点，通常含不严格的反向重复序列，有时在IS元件两端还有转座过程中产生的特征性靶位点的复制（正向重复）。根据一个简单的命名系统可以给新发现的IS元件命名：属名的第一个或前两个字母+种名的前两个字母+独特的识别数字。原则上，90%以上核苷酸序列都相同的IS元件起相同的名字。这些名称可以保存在IS-Finder数据库中（http://www-is.biotoul.fr/），这是一个专门对IS元件进行登记、分类和描述的网站。

复制起点。一些寡核苷酸片段趋于在微生物复制起点周围不均衡分布[9]，根据这些片段可以推测出复制起点。另外还可以根据起点附近经常出现的基因来判断复制起点。

并系同源基因家族（paralogous gene family）。并系同源基因指在一个物种内复制的基因。由于基因的不断复制反映了物种在其生存小环境中的生物活动[10,11]，识别这些并系同源基因非常重要。把基因归纳到并系同源基因家族，就增加了每个基因指定的可靠性。识别和注释并系同源家族的方法比较简单，主要是对微生物的所有蛋白质用相当严谨的参数进行搜索，然后对搜索结果加以审查。但是，把蛋白质划分到并系同源家族不能仅仅用一个吻合标准。并系同源基因间的相似性是基因在漫长进化过程中复制的结果，如何解释这些结果就不可避免因人而异，况且也因微生物种类和基因家族的不同而不同。

双序列蛋白搜索。先用BLAST算法在非冗余性公共蛋白数据库中，对预测编码区进行搜索，搜索结果收集到微型数据库中，然后用PRAZE软件在DNA水平上把预测编码区的延长片段与微型数据库中的条目进行序列比对。PRAZE是应用Smith-Waterman算法进行序列比对的程序，它能跨越序列间隔区，还能切换到其他读框进行序列比对，因此，PRAZE对识别移码突变非常有效。

ORF的起始/终止和管理。目前预测翻译起始点的准确率达75%，注释员通过图形化用户界面检查Glimmer结果，把吻合条目与很多直系同源蛋白比较，并分析其上游基因，以便最准确确定翻译起始点。将含有移码突变的区域挑出，再用其他测序方法验证。在典型细菌鸟枪测序计划中，能发现和修正约200个移码突变。

结构RNA。tRNAscan识别转运RNA。人工识别核糖体RNA和其他结构RNA。

表2中的数据说明，自动注释是成功的，表中的第一部分评估了自动指定EC号和遗传名的可靠性，将自动注释结果与人工注释结果进行比较。衡量标准是自动注释的特异性（specificity），即它的准确性。在对5个基因组的所有基因指定遗传名和EC号时，自动注释的平均特异性分别为50.4%和49.9%。敏感度（sensitivity）指正确注释的基因数占所有应注释基因数的百分比，对所有应注释的基因，自动指定遗传名和EC号的敏感度分别为75.3%和51.9%。

表2中在指定基因俗名时，还比较了自动方法与人工方法的差异，结果显示平均47.5%的自动指定俗名被人工改动(提高)。不能像衡量自动指定的基因遗传名和EC

表 2　自动注释与人工注释的比较

	A 部分											
	炭疽芽孢杆菌		猪布鲁氏杆菌		无乳链球菌		恶臭假单胞菌		*Shewanella oneidensis*		总计	
	填充	空缺	填充	空缺	填充	空缺	填充	空缺	填充	空缺	填充	空缺
遗传名												
正确	580	1964	601	1109	440	749	668	1912	815	1321	3104	7055
假阳性	817	200	421	147	328	67	883	410	611	195	3060	1019
假阴性	200	529	147	333	67	263	410	588	195	388	1019	2101
灵敏度/特异性	74.4/41.5	78.8/90.8	80.3/58.8	76.9/88.3	86.8/57.3	74/91.8	62/43.1	76.5/82.3	80.7/57.2	77.3/87.1	75.3/50.4	77.1/87.4
EC 号												
正确	331	2488	251	1516	140	1086	382	2798	358	2065	1462	9953
假阳性	406	336	254	257	171	187	374	319	262	257	1467	1356
假阴性	336	386	257	240	187	165	319	343	257	233	1356	1367
灵敏度/特异性	49.6/44.9	86.6/88.1	49.4/49.7	86.3/85.5	42.8/45	86.8/85.3	54.5/50.5	89.1/89.8	58.2/57.7	89.9/88.9	51.9/49.9	87.9/88

	B 部分											
	炭疽芽孢杆菌		猪布鲁氏杆菌		无乳链球菌		恶臭假单胞菌		*Shewanella oneidensis*		总计	
俗名												
预测	3561		2278		1584		3873		2942		14 238	
修正	1,678	47.1%	970	42.6%	521	32.9%	2492	64.3%	1097	37.3%	6758	47.5%

注：表 2 中列举了炭疽芽孢杆菌（*Bacillus anthracis*）、猪布鲁氏杆菌（*Brucella suis*）、恶臭假单胞菌（*Pseudomonas putida*）、*Shewanella oneidensis* 以及无乳链球菌（*Streptococcus agalactiae*）的注释结果。A 部分是自动指定遗传名和 EC 号的结果，自动预测要么给基因指定一个遗传名或 EC 号（填充，filled），要么在该序列区显示空白（空缺，null）。计算机指定完成后，注释员对每个基因的相关信息（如 BLAST 和 HMM 搜索结果）加以审查，判断这些指定是否正确，是否需要改进，正确的指定次数、假阳性和假阴性出现的次数都统计出来。灵敏度值由"正确次数/（正确次数 + 假阴性次数）"计算得出，特异性由"正确次数/（正确次数 + 假阳性次数）"计算得出。B 部分比较了自动和人工指定俗名的差别，表中显示修正过的基因是由忽略字母大小写区别的字符串吻合算法（case-insensitive string-matching algorithm）统计出的。

号那样，无法用简单的字符串比较（string comparison）决定自动指定俗名是否正确。因为在很多情况下，自动指定和人工指定的区别仅仅在于一些语法上的小出入，并不改变基因名称的含义，换句话说，自动指定的名称并不是错误百出，只是它们本可以更精确一些。这些改动有时是为了消除歧义（如将“可能的芽孢萌发蛋白 C”改为“芽孢萌发蛋白 GerPC”），有时是为了把俗名下的信息划分到别的数据库区域中［如把“过氧化物歧化酶（peroxide dismutase）［EC 1.15.1.1］”改为“超氧化物歧化酶（superoxide dismutase)”］。有些蛋白质与公共档案中的蛋白质条目吻合，实验分离这些基因时带有的一些信息（如“RXA00030”，假定蛋白），这些名字就改为“保守假定蛋白”。有些情况只是为了遵守命名标准［如把“ABC 转运蛋白‘寡’多肽（peptide ABC transporter)”改为“ABC 转运蛋白多肽（oligopeptide ABC transporter），ATP 结合蛋白”］。总之，表 2 中的数据显示，自动指定的成功率相对较低，自动指定的灵敏度和特异性统计数字，以及需要人工修正俗名的数目，进一步强调人工整理对产生高质量注释细菌基因组的重要性。

TIGR 制定给基因指定功能的 SOP，包括确定俗名、EC 号、遗传名以及作用类型的方法。管理员在指定基因功能时，使用一个图形用户界面 Manatee（图 2）读取数据，该界面可使管理员轻松地读取每个预测编码区的所有数据。主信息页显示，识别信息并总结各个搜索软件的运行结果。用户还能看到 TIGRFAM 的打分结果，还可通过链接进入内部和外部网页，这些网页提供任何给定模型的详尽描述，以及预测蛋白与建立该模型所需的所有蛋白质的多序列比对（multiple alignment）。预测蛋白中找到的 InterPro[6]模体（motif），也与对应序列片段列在一起，还可以通过链接获取 InterPro 文件。

Manatee 还显示总结 BER 搜索结果的表，并通过链接把用户引导到吻合蛋白的原始数据库条目，在这些数据库中，注释员可以得到吻合蛋白的相关信息，如活性位点、跨膜区域以及脱氧核糖核酸（deoxyribonucleic acid，DNA）结合位点。注释员要核对这些信息，看它们是通过实验确定，还是通过序列相似性推断来的，并利用这些信息评估预测蛋白是否与吻合蛋白具有相同的模体、结构域（domain）以及功能。

信息页还显示预测编码区的物理特性，例如基因的起始和终止位点（基因坐标）、基因及其蛋白质长度和分子质量以及等电点，注释员可以通过链接查看 DNA 和蛋白质的序列、跨膜区以及基因组中该基因周围区域的图形显示。

将各方面信息精炼后，给基因准确指定功能是非常复杂的任务。注释员们尽其所能给每个基因提供可靠信息，以免从序列相似性中推导得太多。这就要在基因命名时持保守态度，因此命名系统中基因名称的特异性就反映了注释员对某一基因名称的信任程度。如果有多方面证据说明一种蛋白质具有某项特殊功能，包括隐式马可夫模型（hidden Markov model，HMM)[7]的吻合、多组全长双序列比对达 30％以上的吻合，以及保守 Interpro 模体（假如适用的话），那么就给蛋白质起一个有详尽描述的名字，并给基因指定一个名字，如“核糖 ABC 转运体，渗透酶蛋白（*rbsC*)”，这样的名字反映了高信任度。但是，如果搜索结果显示这个转运蛋白底物可能是多种糖类，注释员就会给蛋白质起一个较广义的名字，如“糖 ABC 转运体，渗透酶蛋白”，也不给基因指定名字；如果这个转运蛋白的转运底物还不清楚，就命名为“ABC 转运体，渗透酶蛋白”；最后，如果连转运蛋白的类型都不清楚，就命名为“转运体，假定蛋白”。有时，只明确

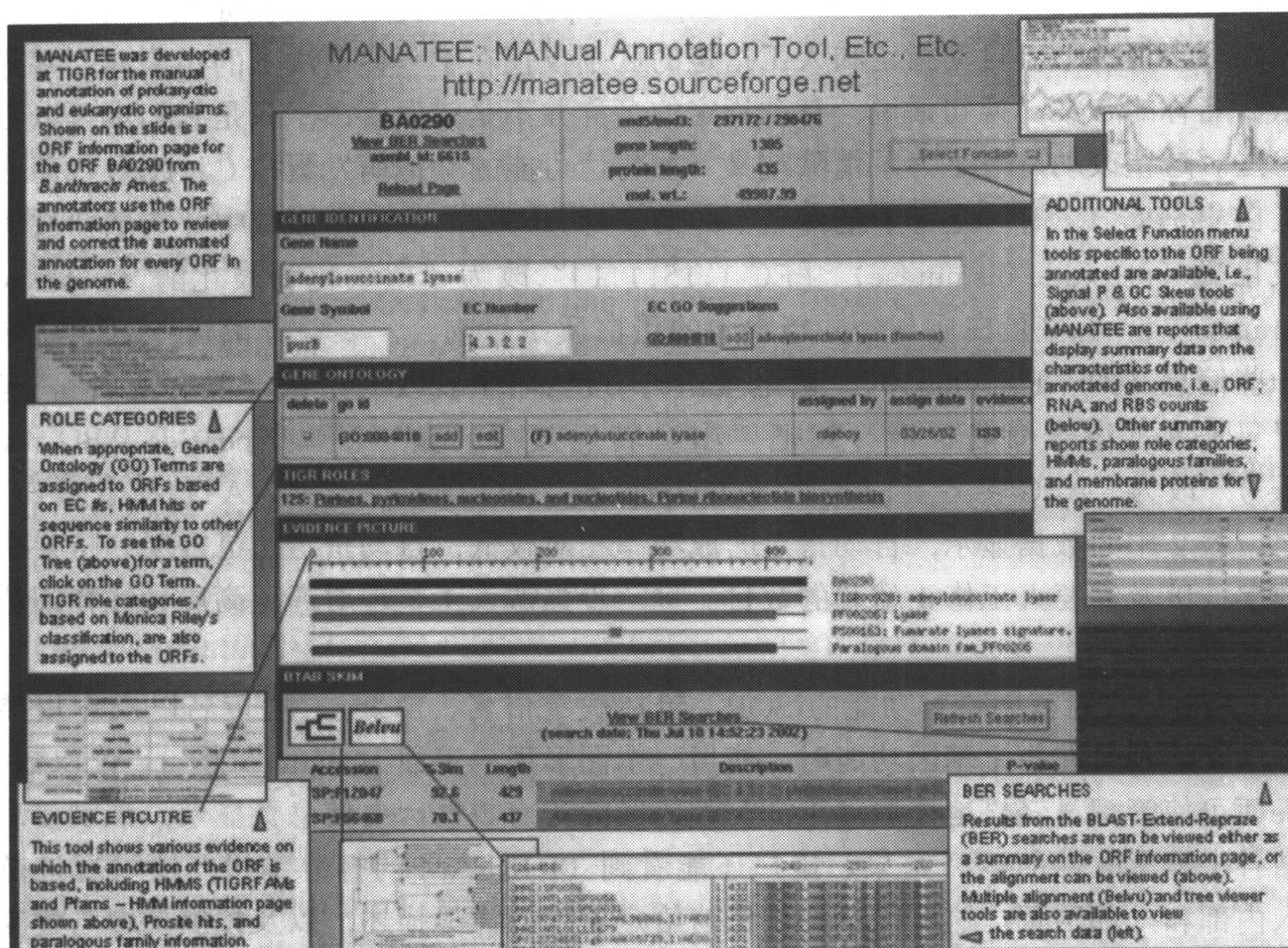


图 2　Manatee 注释系统。该系统以内部浏览器 Web 形式和点击界面构建。图中显示 Manatee 界面的几个模块，可使生物学家们能快速识别基因，从而产生高质量功能指定。一个注释项目可从几个不同方面加以评估，用户可从染色体、基因本体论、并系同源蛋白或 InterPro 结构域等各方面查看数据。Manatee 界面还支持以不同特性选择基因，如基因酶学委员会编号、基因符号以及 HMM 吻合或 COG 存取号（accession）。用户还可得到像移码突变报告或重复序列分析这样的总结信息。Manatee 还提供一些其他方面的总结报告，如注释工作进展、基因组整体基因含量（genetic content）、注释员联系方式以及全基因组图形直观图。系统要求：Manatee 由 Perl 语言编写，在 Web 服务器（如 Apache）上运行，Manatee 还要求有至少一个 MySQL（或 Sybase）项目数据库以及相关搜索文件。项目数据库所用数据模型和框架是由 TIGR 为了存储真核或原核生物数据而开发。在 manatee. sourceforge. net 中列举了一些数据库、使用手册和源代码。

知道蛋白质属某个特定蛋白质家族，在这种情况下，就只给该蛋白质用它的家族名，只有这样，才能在蛋白质俗名中尽可能多地包括可以信任的信息。

基因家族数据库 TIGRFAM

在给新发现基因指定假定功能时，序列相似性是最常用的方法，另外还有以序列为根据预测蛋白功能的方法，如蛋白模体搜索和特殊组成算法（如信号肽和跨膜区的计算）加以补充。但是，目前大多数的功能指定都是“转移式”指定（“transitive” assignment），都是将未知基因与公共蛋白档案中的基因进行双序列比对得到。名字的指定通常反映对每个基因的谨慎解释，但是，注释过程还可能包括许多错误指定，例如，功能指定只是某个蛋白质与某一查询序列（query sequence）吻合，但与它们的功能不见得

一致。注释员可以对基因组中的某些基因有深厚造诣，但并不是对所有基因都很熟，所以只能根据他们所擅长的领域给予正确注释。由于这样那样的种种原因，微生物全基因组中的数据存在许多问题，当查询序列本身含有模棱两可的指定功能时，由此再给下级序列做准确转移式指定几乎不可能。

把结构基因组织到直系同源家族［ortholog family，（功能相关的基因群体）］，有助于解决转移式指定带来的问题。转移式指定失败的原因在于没有单独的指定标准，如BLAST概率，可以用来给所有细菌蛋白作准确双序列指定。在人工组织直系同源家族时，可依照一定标准把单个基因组织成家族，从而就产生了新成员加入这个家族时应符合的标准。家族成员的多序列比对可提供基因相似程度及其功能域相似程度的范围，这对正确识别基因起关键作用。与双序列相似性搜索方法相比，多种细菌基因组计划产生的冗余性更有利于预测基因功能，而不是起反作用，这是因为生物相关性使基因比对更具有统计学意义。基因家族的管理要求，当把基因划分到直系同源家族时，再次评估每个基因的功能，由转移式注释错误指定的功能通常在这一过程中纠正。

直系同源家族的另一个作用是用它们来构建HMM和搜索新测序基因组序列，然后用HMM来搜索、评估暂未定名基因，并把它们划分到相应的直系同源家族中。TIGR的直系同源蛋白数据库[5]（即TIGRFAM）已经构建成HMM，它明显比经典序列相似性搜索灵敏得多。TIGRFAM的内容在基因注释中也有重要价值，这不仅因为检索的灵敏性，而且还因为数据库条目本身可从多方面注解。每个TIGRFAM条目都有一个家族阀值（cutoff score），在做基因指定时，根据这个阀值能可靠地给未定名新基因预测功能。TIGRFAM的条目说明和引用文献可以告诉我们，搜索到吻合区域后应该如何解释这些结果。确认的基因和例外的基因都列出来，注释常犯的错误也可能指出来，TIGRFAM的条目还会有一些基因别名的总结，以澄清混淆。

在注释基因时，既用TIGRFAM，也用蛋白家族Pfam[12]。但这两个数据库的重点却有很大区别。Pfam是为搜索大规模原核和真核蛋白序列而设计的，它的设计原则是用一个广泛的模型尽可能覆盖多的同源序列，绝不允许交叠（overlapping），如果某蛋白序列中有两个区域的计算值分别大于两个不同的Pfam HMM阀值，那么这两个区域中绝对不能交叠相同的氨基酸残基序列。这种彻底性和试图在比对过程中各种序列的处理方法使这些HMM的灵敏度极高，这些HMM很管用，它们能鉴别出那些不容易在BLAST检索结果中标注的区域。

然而，Pfam模型的广泛性既是优点也是缺点，一个Pfam模型可能产生功能极其多样化的家族，以致用该模型的名字命名蛋白质会广义到没有实质意义的程度。大多数TIGRFAM模型与一个或几个Pfam模型相交叉，但是却以另外方式处理同样的蛋白质。总之，TIGRFAM的吻合区域较长，每个家族包括较少成员，一个Pfam家族会分割为5个或更多TIGRFAM模型。另一方面，一个TIGRFAM模型可能代表一个具多结构域的长蛋白，而每个结构域都由单独一个Pfam HMM模型界定。

微生物基因组注释必须把TIGRFAM模型的范围限定在有相同或相似功能的蛋白质家族中，并用全长序列作同源性比对，其目的是给整个蛋白质起个有功能描述的名字，而不是给蛋白质列表清清楚楚标出各个结构域的边界和活性位点。细菌基因组中，约有25%指定了功能的蛋白隶属某个TIGRFAM基因家族。

微生物资源大全

目前，很多网址可以提供微生物全基因组的数据信息。表 3 列举了一些这样的网址。此外，TIGR 还创建了微生物资源大全数据库（Comprehensive Microbial Resource, CMR），它是所有完成了的微生物基因组的总库，用户可以从 www. tigr. org/CMR 访问 CMR。为了能横跨所有细菌检索生物信息，如功能、作用类型、蛋白质相似性、疏水性或遗传符号，CMR 数据库仔细解析了各个基因组中的相关数据类型并把它们分别存储。这就使跨越基因组的发现更简便，也更有意义。例如，可以从所有完成的细菌基因组中获取相同生物功能的基因（如"显示所有与氨基酸合成有关的基因"），还可以用多种性质（如分类、与其他蛋白的相似性、革兰氏染色或染色体拓扑学）的限制进行这样的检索，"显示有 5 个以上跨膜区域和分子质量在 36 到 51 kDa 之间所有转运蛋白"。

表 3 微生物基因组资源

蛋白直系同源群簇（Cluster of Orthologous Group, COG）。全基因组蛋白的种系分类，每个 COG 由起源同一保守域单个蛋白质或并系同源蛋白群组成。http://www.ncbi.nlm.nih.gov/COG/

已测序微生物基因组的 DNA 结构分析。一种显示大片段 DNA 结构特征的方法。http://www.cbs.dtu.dk/services/GenomeAtlas

欧洲生物信息中心（European Bioinformatics Institute, EBI）。管理生物数据库，包括核酸、蛋白质序列和大分子结构。http://www.ebi.ac.uk

微生物基因组信息代理（Genome Information Broker, GIB）。从国际核酸序列数据库（International Nucleotide Sequence Database, DDBJ/EMBL/GenBank）中编辑出的微生物全基因组序列。http://gib.genes.nig.ac.jp/

高质量自动和人工注释的微生物蛋白质组（High-quality Automated and Manual Annotation of Microbial Proteomes, HAMAP）。这项计划旨在对微生物基因组测序所产生的相当一部分蛋白质序列进行自动注释。http://us.expasy.org/sprot/hamap

京都基因和基因组百科全书（Kyoto Encyclopedia of Genes and Genomes, KEGG）。根据相互作用的分子或基因组成的信息途径，将分子生物学和细胞生物学的现有知识计算机化，并为基因组测序计划产生的基因目录（gene catalog）提供链接。http://www.genome.ad.jp/kegg/kegg2.html

微生物基因组数据库（Microbial Genome Database, MBGD）。从不同方面为比较基因组学研究提供方便，如识别直系同源序列、收集并系同源序列和蛋白模体分析。http://mbgd.genome.ad.jp

国立生物工程信息中心（National Center for Biotechnology Information, NCBI）。作为国家分子生物学信息资源中心，NCBI 创建公共数据库，开展计算生物学研究，开发分析基因组数据的软件，并传播生物医学信息。http://www.ncbi.nlm.nih.gov/

鲍森转运蛋白页（Paulsen Transporter Page）。比较基因组间的膜转运系统。www.membranetransport.org

蛋白提取、描述和分析工具（Protein Extraction, Description, and ANalysis Tool, PEDANT）。为所有已测序基因组提供基因组分析和注释。http://pedant.gsf.de

基因组渠道（The Genome Channel）。分析预测的基因和蛋白模型，包括计算机注释的基因组。http://compbio.ornl.gov/channel/

那有什么？（What Is There? WIT）。管理基因的功能性指定，开发代谢模型。http://wit.mcs.anl.gov/WIT2/

CMR 中的数据可以在几个水平显示，如在基因水平，用户可查看单个基因的基本指定俗名、自动注释程序所指定的俗名、遗传符号、EC 号、等电点、分子质量、DNA

序列以及所编码的蛋白质。用户还可以得到疏水性图、与其他网址（如 Swiss-Port）的链接、二级结构以及第三位点 GC 偏衡（third-position GC skew）。用户研究某基因时，也许还要与其他微生物具有类似功能的基因相比较，在提供这项服务时，系统会清楚地列出将这些基因联系在一起的证据。这些证据可能来自 TIGRFAM 或 Pfam 蛋白家族，也可能来自蛋白直系同源群簇（Clusters of Orthologous Group，COG），或序列相似性，或共同 EC 号，或共同作用类型。

另一个 CMR 数据显示水平是针对单个微生物基因组，这方面的信息包括把基因在一段染色体上的线性排列图示或把整个染色体显示为完整的环状。有些 CMR 网页还总结了密码子使用频率、GC 图（GC plot）、计算机再造的二维蛋白胶（2D gels），有些还列表总结基因的平均长度、编码区数目等诸如此类的信息。CMR 还包括微生物基因组间相互比较的信息，CMR 网页将每个基因与许多数据类型相连，如疏水性、与蛋白质模体的吻合程度、作用类型以及 EC 号，因此，可以用作进行物种间的大规模比较，和其他此类服务，包括比较不同物种间的蛋白质相似性以及对它们的全基因组进行序列比对。

其他服务

TIGR 还有其他方面的举动会引起参与微生物注释科学家们的兴趣，我的课题组为测序中心免费提供原核生物自动注释系统——注释引擎（annotation engine），这项服务为细菌基因组序列进行全套注释，注释结果随后返回到测序中心。该结果包括可读框和 RNA 起始点和终止点，蛋白质俗名、基因符号、EC 号、TIGR 作用类型，以及基因组本体论命名结果，BER 搜索结果，HMM，InterPro，信号肽分析以及跨膜区域。这些信息存储在 MySQL 数据库中，该数据库可以与 TIGR 人工注释工具 Manatee 一起使用。

注释引擎有多项功能，注释工作从它开始，这有利于提高数据类型的一致性。它也鼓励测序中心采用 TIGR 的 SOP，使注释标准统一化，也使注释结果直接输入 CMR 数据管理系统，以便在 CMR 网页上发布（如果测序中心同意），也让测序中心在自己的局域网上发布前期注释结果，这样就以较小生物信息学资源投资对他们的测序工作表示了认可。由于注释工作分散，参与工作的人力就要合理分配，工作规模适当分割，这样就提高了我们的测序能力。

另外，TIGR 还举办微生物注释训练班，提供大量可能对基因组注释很有价值的软件。欲知详情，请访问 TIGR 的网址 www.tigr.org。

致谢

本工作由美国能源部生物与环境研究办公室（US Department of Energy，Office of Biological and Environmental Research）资助，合作协议号 DE-FC02-95ER61962，8 号修正案。

（许朝晖，喻晓辉　译）

参考文献

1. Riley M. Functions of the gene products of *Escherichia coli*. Microbiol Rev 1993; 57:862–952.
2. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25–29.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol 1990; 215:403–410.
4. Waterman M. General methods of sequence comparison. Bull Math Biol 1984; 46:473–500.
5. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res 2003; 31:371–373.
6. Mulder NJ, Apweiler R, Attwood TK, et al. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003; 31:315–318.
7. Eddy SR. Hidden Markov models. Curr Opin Struct Biol 1996; 6:361–365.
8. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with Glimmer. Nucleic Acids Res 1999; 27:4636–4641.
9. Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF. Skewed oligomers and origins of replication. Gene 1998; 217:57–67.
10. Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995; 270:397–403.
11. Tomb JF, White O, Kerlavage AR, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997; 388:539–547.
12. Sonnhammer ELL, Eddy SR, Birney E, et al. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998; 26:320–322.

4

生物信息学与微生物致病作用

Fiona S. L. Brinkman and Joanna L. Fueyo

传染病是导致40岁以下人死亡的元凶。如果把由感染引起的一些疾病（如呼吸道疾病、消化道疾病等）包括进去，它就是导致全球人类死亡的元凶[1]。耐药性细菌的不断出现又一直威胁着人类健康，同时，人类还面临着传染性生物武器、新病原以及再现（reemerging）病原的威胁。针对这些问题，测定了越来越多病原的全基因组序列，以便了解这些病原的遗传信息组成，最终获得控制的方法。用全基因组方法（如微阵列分析、活体表达技术和蛋白质组方法等）研究病原微生物已经越来越普遍，该方法的基本前提是：通过基因组高度平行化方式研究微生物基因，可加快发现微生物致病机制，并缩短鉴定抗感染药物靶位点和开发疫苗所需的时间。随着基因组数据的快速增加，急需能将这些数据进行归类和分析的生物信息学方法和工具。对于任何研究传染病跨学科的微生物实验室，掌握这些方法和工具越来越有必要。

因此，本章对研究病原微生物基因组和微生物致病性的计算方法作一个概述，必须指出，分析病原微生物基因组及其相关数据的计算方法十分有限，特别是在发现新的毒力基因方面，况且，围绕毒力的定义和毒力因子的组成还有不少争议，这就使分析工作更加复杂化。由于这些争议，就难以确定这些计算方法在鉴定毒力基因或毒力途径过程中的准确性。尽管如此，如本章所述，在这一领域还是有一些成功的例子，但是，很多分析方法和分析工具还有待进一步完善，生物信息学研究本身以及生物信息学对研究重要传染病均有巨大的潜力。

基本思路

本书有关章节提到的很多常用生物信息学方法，也可用于研究病原微生物。为简明起见，本章只扼要描述其中部分方法，重点强调那些更适于研究病原和致病性的成熟分析方法，并选择性阐述研究病原的一些生物信息学方法。尽管随着生物信息学朝着群体基因组学方向的发展，流行病学方法可能会整合到本章所涉及的方法中，但在这里不单独阐述，此外，种系发生和相关进化分析为认识毒力进化和致病机制提供了重要线索（参见文献［2～5]），本章也不一一例举，本文侧重用于研究细菌病原的方法，主要是最近发展的方法。读者可根据本章提及的网站和引用的文献了解更多的信息，这些引用文献通常会提供更为详细的有关每种方法的设定、不足、合理应用以及解释分析结果的建议，了解这些后，就可以避免夸大和低估计算分析方法的作用。

病原生物信息学和应用计算方法研究微生物致病作用，是一个相对较新的研究领域，其发展十分迅速[6~8]，本章提到的只是分析病原和微生物致病机制的生物信息学方法的基本框架，相信这些方法会越来越完善。

通过同源序列分析或蛋白模体（motif）分析鉴定毒力因子：对毒力基因数据库和疾病特异本体论的需求

最常用分析病原基因组的生物信息学方法是鉴定与已知毒力基因同源的基因。有趣的是，迄今还没有用于这种分析的完善的综合分析工具，常用的序列相似性分析方法，包括BLAST，PSI—BLAST和FASTA[9-11]，而Smith-Waterman方法比较少用，该方法相对更敏感，但速度慢[12]。随着更多毒力因子的蛋白结构被推断出来，如VAST（http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml）的蛋白结构分析方法可能会用得越来越多。用户最好从那些应用BLAST及其相关方法过程中，从指出了应注意事项的教科书[13]或文章[14]中获得有关这些方法的信息。当然，所有这些数据库搜索方法的应用，还有赖于用户对数据库中哪些基因是毒力基因，或所研究毒力基因数据库发展状况的了解程度。但是，目前几乎还没有公开的毒力基因数据库，最常见的是人工或半人工方法评定所分析基因与选择的毒力因子序列相似性，或者构建一个内部毒力基因数据库。

缺乏毒力基因数据库的原因之一是毒力基因的定义还不确定，争议很大，传统定义致病生物的柯赫法则（Koch's Postulate），已修正为定义毒力因子的“分子柯赫法则”。随着对微生物—微生物基因—微生物所处环境间的相互作用和寄主对疾病遗传敏感性的深入了解，毒力因子的概念也不断被进一步修改[6]。尽管如此，还是构建了一些小范围的毒力因子数据库，如BacBix毒力因子数据库（http://www.jenner.ac.uk/BacBix3/Welcomehomepage.htm），该数据库来源于PRINTS数据库[15]。如果采用适当的搜索，Swiss-Prot以及其他人工或半人工建立的基因数据库，也可以作为已知的毒力基因资源，这些资源有待进一步完善。主要基因组中心或研究分子致病作用的实验室，已经收集了或正在收集感兴趣的毒力基因数据库，有些毒力基因数据库将会公开。

对未来开发毒力因子数据库最有用的途径之一是定义毒力相关属性，以及确立它们之间关系的本体论（ontology，一套确定的词汇）。由于毒力的复杂属性（依赖于许多因素），应该把有关毒力不同水平的基因功能，以及基因与环境间的相互作用一并考虑。只有当毒力基因数据库中的每个基因按照以上本体论进行归类，并详细描述某一毒力因子在致病过程中发挥作用应具备的条件，该数据库才具有价值。这样，对毒力因子同源物进行评价时，才能更准确地预测它们的真正作用，或者在毒力方面发挥作用应具备的条件。

其他寻找基因组中可能毒力因子的方法，就是检查被分析基因中是否有毒力因子的特有模体或蛋白质结构域。所采用的方法和利用的资源，如InterPro（提供PROSITE，Pfam，PRINTS，ProDom，SMART和TIGRFAMs的交互查询）在文献中有更详细的描述[16]，这些资源大多不是用于鉴定与致病作用有关的基因，因此，在应用过程中应根据用户的兴趣进行特定分析。

值得注意的是，如果有了适当的数据库或对已知靶位点有彻底的了解，就很容易转变为寻找药物作用靶位点和疫苗的方法，这些方法的思路是：假如某种抗微生物药物能够结合到某特定蛋白上，那么，在另外一种类似微生物中，该蛋白的同源物也可能适合

作为药物靶位点。因此，为了使这些方法能够在正确预测靶位点时有合理的准确性，需要建立适当的本体论。

不管是否能发现毒力因子或治疗靶位点，在序列相似性分析和模体分析过程中，应该谨慎行事，原因是还很少有关这些方法在发现新毒力基因方面的准确性评价，而且，某一基因可能在某一病原中是毒力因子，而在其他病原或不同环境中未必就是。但是，如下所述的一些研究表明，这些方法可以帮助挑选出那些值得进一步研究的基因，因此，这些方法还普遍采用，希望将来对这些方法进一步程序化和准确化。随着对越来越多基因组进行的分析，这些方法在分析不同基因家族的成功率会越来越高。

基因组比较法寻找致病岛及其相关序列

20世纪80年代后期，自致病岛（pathogenicity island，PAI）首次在泌尿道致病性大肠杆菌基因组中被发现和命名[17]以来，研究微生物致病作用的很多实验室对致病岛进行了深入研究。致病岛涉及两个有趣的现象：细菌的致病作用和基因的水平转移(horizontal gene transfer，HGT)。很多与致病有关的基因都是成串排列[18,19]，而且越来越多的证据显示，这些基因簇起源于基因的水平转移[20,21]。致病岛的概念可以扩展到其他遗传组分，这些遗传组分具有致病岛结构特征，但与毒力无关[22]，因此，这些组分统称基因组岛（genomic island)，如次级代谢岛、抗生素抗性岛、分泌岛。所有这些遗传组分可为生物提供功能优势，致病岛中的致病基因能促使微生物成功地感染人体。

基因组岛的特征包括侧翼重复序列、移动基因（整合酶基因、转座酶基因)、近侧转运核糖核酸和异常G+C百分比[23]。转运核糖核酸是噬菌体的整合位点[24,25]，也可能是移动遗传组分的整合位点，这些移动遗传组分整合后就成为岛。G+C百分比和种特异性DNA特征有助于岛的鉴定[26,27]，因为岛内特征通常与基因组的其他部分明显不同。

目前只有IslandPath一种计算工具综合多个特征去搜索岛，它是一种网上工具，可列举全基因组中与岛相关的特征[28]，显示预测可读框的G+C百分比、基因簇的二核苷酸偏向性（一项独立的基因组组分指标)、已知或可能的移动基因以及转运核糖核酸。在一个压缩图表中，这些特征分别由不同符号代替（图1；http://www.pathogenomics.sfu.ca/islandpath/)。该工具可以根据用户需要而设定不同G+C百分比阈值，其界面与美国生物技术中心（National Center for Biotechnology Information，NCBI）的有关基因注释和分析网页相链接。IslandPath旨在快速预览某一生物的基因组，找出可能感兴趣的岛以便作进一步计算和实验分析，第一版IslandPath可以分析所有目前公布的细菌和古生菌的全基因组（http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html)，其数据库也会经常更新。

应用这种基于DNA组分的方法时要注意，一些DNA信号序列组成不能很好地体现基因的水平转移特征[29]。然而，IslandPath除了对多种DNA信号序列进行分析外，还会考虑其他一些与已知致病岛相关的特征，而且，人们还在考虑把已知毒力基因的同源性信息也添加进去。当然，如果毒力基因不通过水平转移获得，或者其DNA组成特征与基因组的其他部分相似，那么IslandPath和其他DNA序列组成的分析方法，就不

能有效地识别这些致病岛和毒力基因，所以，它们只能筛选出一部分可能的毒力因子。有趣的是，基因组中绝大多数典型毒力基因，都位于 DNA 信号特征不同区域内（F. S. L. Brinkman，2003，未发表数据；有关例子可见 IslandPath 在线帮助文件）。

通过首先发现致病岛而寻找毒力基因的方法，在微生物致病作用研究中取得了显著的成功[23]。希望这种方法也能帮助检测病原菌中其他感兴趣的基因，如抗生素抗性基因，由于这些基因通常以岛的形式存在，因此，也具有一些可以用计算方法分析的共同特征[22]。

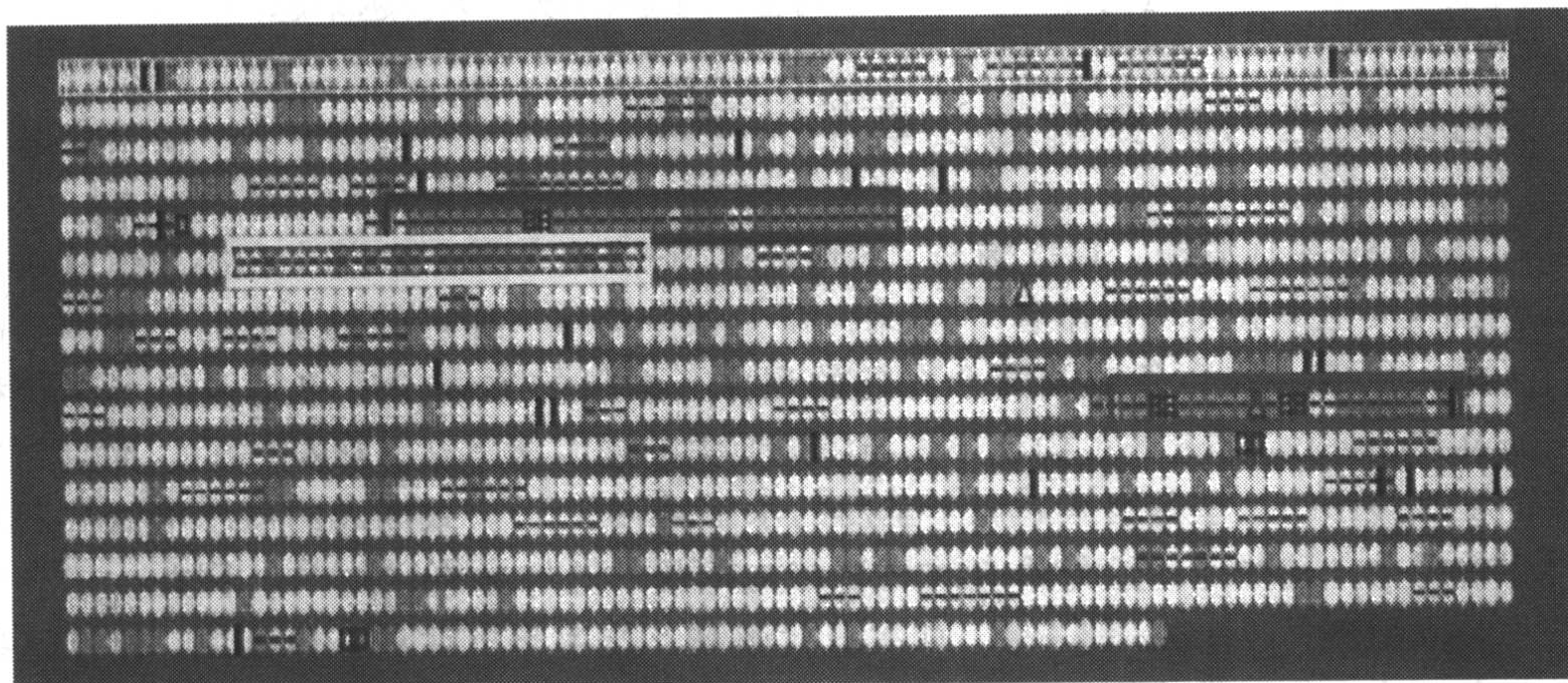


图 1 IslandPath 对幽门螺杆菌菌株 26695 全基因组进行分析的输出结果，列出了所有已知致病岛和可能的致病岛。每个圈代表一个预测编码蛋白的可读框。圈的颜色表明 G+C 百分比（黄色代表高于选定的高阈值，粉红色代表低于选定的低阈值，绿色为两者之间）。圈中的横线代表二核苷酸偏向性[28]，竖杆表示转运核糖核酸和核糖体核糖核酸（黑色为转运核糖核酸，紫色为核糖体核糖核酸，深蓝色为转运核糖核酸和核糖体核糖核酸）。黑色方块为已知的或可能的转座酶基因。黑色三角表示已知的或可能的整合酶基因。具有几项这样特征的区域可能为基因组岛。该菌株中 3 个已知的或可能的岛以彩色框表示：黄框，CAG 致病岛；蓝框，含有 *virB* 基因的同源序列但在幽门螺杆菌菌株 J99 中不存在的区域；红框内的基因在菌株 J99 和 26995 中不尽相同。请注意，有二核苷酸偏向性的大片段区域（长横线）与已知的或可能的基因组岛有很好的相关性。有关 IslandPath 的使用在网上以沙门氏菌（*Salmonella*）为例另有介绍（http://www.pathogenomics.sfu.ca/islandpath/）。由 IslandPath 新近找出已知的或可能的岛均被标示。（另见文前彩色插图 4-1）

比较基因组法鉴定致病岛和毒力特异序列

用比较基因组法对病原进行生物信息学分析，基于一个简单的前提，即两种微生物致病力不同或相似，在其基因组序列中反映出来。近年来，开发了一些比较基因组工具，用这些工具可以比较整个微生物基因组，从而鉴定某些基因的存在是否与某一特定致病表型具有相关性。主要有两种方法用于这种全基因组比较，最常用的方法是比较相近微生物的基因组，找到可能与致病性有关的特异性序列。另一种方法正好相反，就是比较那些引起相似感染而基因组差异显著的病原基因组，以找到那些可能与某些毒力表型有关的类似基因。

第一种方法，ACT（Artemis Comparison Tool，ACT，http://www.sanger.ac.uk/

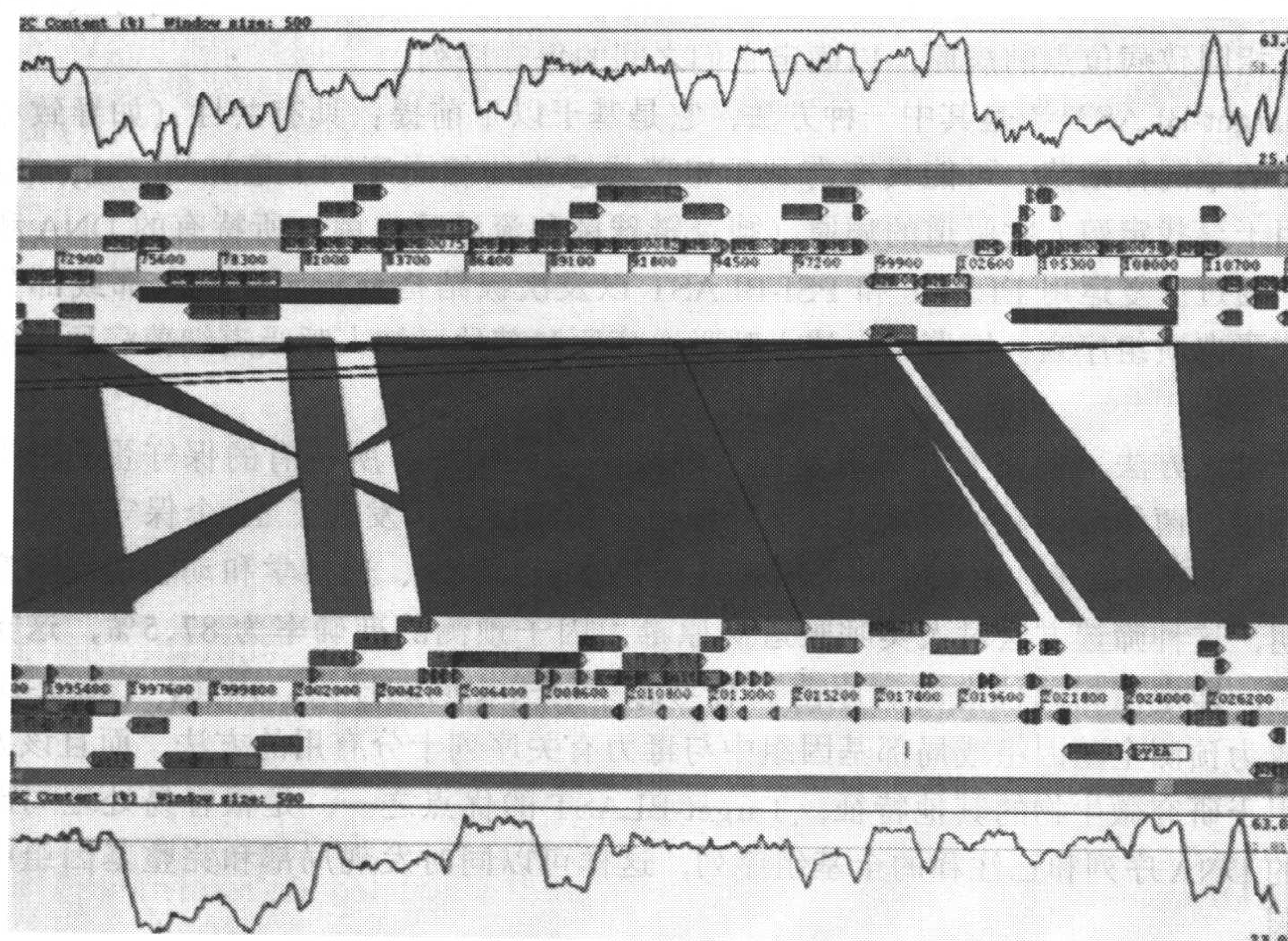


图 2 ACT 分析结果窗口。ACT 可以比较不同病原基因组或病原基因组和非病原基因组。每个红色/粉红色区域相当一个 BLAST 搜索结果，红色代表较好吻合，白色/浅粉红色代表低值吻合。红色/粉红色区域上下方序列分别为参考序列（上方）的正向和反向序列及查询序列（下方）的正向和反向序列；每个方向都有三个可读框。G+C 百分比用斜窗口表示。该图为脑膜炎奈瑟氏球菌菌株 MC58（血清型 B）和菌株 Z2491（血清型 A）。脑膜炎奈瑟氏球菌 A 和 B 血清型有不同的毒力表型。采用这种工具，可将病原菌基因组中与特定疾病表型有关的区域找出来，有利于发现与微生物致病作用有关的序列（本图由英国桑格研究所病原测序部提供）。（另见文前彩色插图 4-2）

Software/Artemis）是常用的全基因组比较法，ACT 的序列比较，通常是对 BLASTN 或 TBLASTX 的搜索结果，用 MSPcrunch[30] 作进一步处理，以提炼出高分值片段。BLAST 搜索必须另外先做，因为 ACT 实际上只是一种浏览工具，图 2 列举了其 Java 环境下的直观输出格式。

该方法的另一组更具体针对肠道微生物基因组，其中包括 Enteric，Menteric，Maj[31]（http://bio.cse.psu.edu）。Enteric 的比较结果是以大肠杆菌作为参照与相关细菌基因组的比较图，每次覆盖 20kb 序列；Menteric 的输出结果是核苷酸水平的多重排列比较，并注明可读框和调控位点，每次覆盖 1kb 序列。Enteric 和 Menteric 都是采用 BLASTZ 算法[32]，Maj 在 Java 环境下编写，结合 Enteric 和 Menteric 的特点，具更多功能，如局部放大功能。

以上工具还可以比较病原与相近非病原，或比较具不同毒力的菌种或菌株。文献报道了许多用这些方法为微生物的致病作用研究提供新研究手段的例子[31,33~38]，值得一提的是人们渐渐认识到微生物种间和菌株间变异的复杂性，比较大肠杆菌菌株 K12 和血清型 O157 的基因组就遇到这些问题[38]，显然，还需要更多工具来应付这些复杂性。

第二种方法采用与第一种方法相反的策略，即比较进化关系远，但可导致相似感染或具有相同致病位点的病原，以确定它们之间的保守序列。

Target-BLAST[39]是其中一种方法，它是基于以下前提：具有共性（如导致相似的疾病）的不同种细菌，可能具有有利于定殖或感染的相同序列。最初的 Target-BLAST 主要用于寻找定殖人呼吸道的病原（肺炎链球菌和流感嗜血菌）所特有的 DNA 和蛋白序列。通过反复运用 BLAST 和 PSI-BLAST 以及次级结构分析，比较全部或部分基因组，或蛋白质组序列，以鉴定上述人呼吸道病原和其他已知人呼吸道细菌病原中的保守序列。

通过该方法，发现了人口鼻咽和下呼吸道感染病原，所特有的保守蛋白序列[39]，在流感嗜血菌和肺炎链球菌以及其他呼吸道感染细菌中，发现了 12 个保守序列，其中 4 个为已知毒力因子，对其余 8 个序列进行了分子生物学、遗传学和动物模型研究。结果表明，这种筛选方法对人类呼吸道病原毒力因子预测的准确率为 87.5%，这支持了该假说，即定殖相同位点的病原菌具有与感染有关相同的序列。因此，Target-BLAST 已经成为预测全基因组或局部基因组中与毒力有关序列十分有用的方法，而且该方法也适合用于研究微生物的其他特征，Target-BLAST 的优点之一，是很容易处理局部且未注释的 DNA 序列和已注释的全基组序列，这样可以同时发现局部和完整基因组中的保守基因。

另一个相关方法为 SEEBUGS[40]，它是采用 FASTA 算法[41]，搜寻已注释的全基因组中已知抗生素靶位点的一致序列。即采用基因组消减法，分析寻找某生物中的一些基因，这些基因在另一类生物中存在，但不在第三类生物中存在，这种方法已用于鉴定细菌病原所共有而在寄主中没有的基因，该方法适合于广谱抗生素靶位点的寻找。值得指出，这种方法只限于已注释的全基因组，也意味着未完成基因组中的基因或蛋白将会被漏查。

这些基于序列相似性的分析方法是很好的前期筛选方法[42]，但有其局限性，它不考虑基因进化速率的差异，期望对未来一致序列的分析能采用更先进的进化分析法，增加探查合适目标序列的能力。

值得一提的是，基于微阵列的基因组比较法，导致分析微阵列数据的新生物信息学工具的开发，微阵列分析在本书其他章节已有详细描述，本章不进行讨论。这些方法已成功用于更深入的微生物致病研究，如通过对地方性和流行性霍乱菌株的比较，发现了与流行性相关的基因[43]。

采用一些分析工具，如 MetCyc 和 Pathway Tools 或 KEGG（Kyoto Encyclopedia of Genes and Genome，KEGG），进行代谢途径比较分析也是很有用的方法，这种方法较适合细胞水平，稍后再对 MetCyc、Pathway Tool 和 KEGG 方法进行讨论。全基因组基因功能和代谢途径的比较分析，有助于鉴定某基因或代谢途径在特定微生物中的功能，从而发现适当药物靶位点和候选疫苗[44,45]。直向同源群簇（Cluster of Orthologous Group，COG）[46]的比较分析，是另一种相关基因功能的分析方法，通过比较 40 种微生物全基因组所编码的蛋白序列，找出其中的直系同源群簇，这些基因很可能具有相似的功能。直系同源基因（orthologous gene）是指在种间分化中产生的相似基因，与之相对的横系同源基因（paralogous gene）则是指由基因复制而产生的相似基因。因此，这种将基因

按直系同源基因簇方式归类的基因组分析方法，有助于发现某生物所特有的基因或某类生物所共有的基因。

随着越来越多基因序列的出现和群体基因组学的发展，所有这些基因组学分析工具，将会在病原生物信息学研究中起重要作用。

预测表面蛋白和分泌性蛋白

由于蛋白质亚细胞定位信息，可为揭示该蛋白质在生物中的功能提供线索，因此，蛋白质亚细胞定位的预测是基因组分析和注释的重要工具。对于病原微生物，预测细胞的表面蛋白十分有意义，因为细胞的表面蛋白具备作为药物靶位点和疫苗的潜力（一般认为不穿过脂质双分子层的药物更容易开发，而疫苗候选蛋白必须和寄主免疫系统接触)。预测分泌性蛋白也十分有价值，分泌性蛋白与被感染寄主细胞的相互作用，可在致病过程中起作用，有很多典型分泌性毒素（外毒素)、蛋白质酶等，均为已知的毒力因子，如霍乱毒素、溶血素、透明质酸酶和免疫球蛋白 A 蛋白酶，一些分泌性蛋白可作为很好的药物靶位点或候选疫苗，如百日咳毒素，该毒素存在第二代白喉—破伤风—百日咳疫苗中。

蛋白质亚细胞定位受其初级结构特征的影响，如信号肽序列和跨膜 α 螺旋，已经开发出一些算法可识别原核和真核微生物蛋白质的某一结构特征（综述见文献 [47])，而且有些方法可以同时鉴定多种结构特征，表 1 列举了预测表面蛋白的一些方法，还列举了膜蛋白拓扑结构的预测方法，也有助于预测膜蛋白细胞表面的暴露序列。

蛋白质亚细胞定位的预测，是一种寻找与寄主相互作用的病原蛋白或是寻找治疗/预防性药物靶点的重要方法，下文提及了成功预测的实例。值得强调的是，每种预测方法的准确性差异很大[47,48]，有些方法缺点明显，例如，最初版本的 PSORT 就不能预测分泌性蛋白，它会强行将某蛋白质归为某一亚细胞位置，这样就错误地预测一些蛋白的亚细胞定位[48]；绝大多数真核生物的预测方法是针对动物、酵母和植物，因此，预测原生动物病原的准确性不高；同样，预测细菌蛋白亚细胞定位的方法，主要针对革兰氏阴性和阳性细菌，因此预测其他类群细菌（如螺旋体和衣原体）的准确性也不高。由于一些预测方法的训练数据很有限，因此，用户最好参考每种方法的原始文献，psort.org 网站提供了一些训练数据可根据用户需求进行特殊分析，另外还提供了一些网上应用方法的链接。

表 1 列举了常用蛋白亚细胞定位预测方法。只有合理使用这些方法，才能达到较好的预测结果，例如，由于亚细胞定位是相当保守的特性[49]，于是就开发出序列相似性分析方法，如 SCL-BLAST（SubCellular Localization-BLAST，SCL-BLAST[48])，然而，由于蛋白质存在结构域，当检索蛋白与目标蛋白长度相似时，SCL-BLAST 的预测准确性会增加。对已知亚细胞定位的 1443 种蛋白进行测试分析结果表明，SCL-BLAST 的准确率达 97%[48]，只有获得更多已知亚细胞定位蛋白，SCL-BLAST 准确率才会进一步提高。另一种常用方法是根据特定亚细胞定位某一蛋白质所特有的模体进行预测，该方法只有采用高度特异模体，才能达到很好的预测结果，如 PROSITE 模体 PS00695（ENT_VIR_OMP_2）是鉴定革兰氏阴性细菌外膜蛋白的高度特异性模体。

表 1 病原基因组中分泌性蛋白、表面蛋白和蛋白质细胞表面序列的预测方法[a]

真核生物和革兰氏阳性细菌细胞表面膜蛋白及其拓扑结构的预测方法

- HMMTOP[73]
- TMHMM[74]
- DAS[75]
- Tmpred[75a]
- SOSUI (Tokyo University of Agriculture and Technoloy)
- TMAP (Karolinska Institut, Sweden)
- TopPRED 2 (Pasteur Institute)
- Janulczyk 和 Rasmussen 开发出用于预测革兰氏阳性细菌细胞壁蛋白的方法[76]

革兰氏阴性细菌外膜表面蛋白或它们拓扑结构的预测方法

- PSORT-B[48]基于辅助载体、BLAST 和模型的外膜蛋白预测方法
- 外膜蛋白 β 链拓扑结构预测[77]
- 外膜蛋白穿膜区拓扑结构预测[78]
- 外膜蛋白预测[79]
- β 折叠桶分析工具 BBF[80]
- 根据蛋白质疏水性预测外膜蛋白[81]
- 根据蛋白质序列组成预测外膜蛋白[82]

分泌性蛋白的预测方法

- 革兰氏阴性细菌分泌性蛋白预测方法 PSORT-B[48]
- 真核生物分泌性蛋白的预测方法 PSORT II[84]
- iPSORT[84]真核生物 N 端分拣信号分类的预测方法
- TargetP[47]真核生物 N 端分拣信号分类的预测方法
- 真核生物和细菌蛋白质信号肽预测方法 SignalP[85]
- SubLoc[86]利用辅助载体对真核或原核生物的蛋白质进行亚细胞定位的预测方法
- NNPSL[87]利用氨基酸的合成对真核或原核生物的蛋白质进行亚细胞定位的预测方法
- 脂蛋白模体分析法 PSORT[88]
- ExProt[89]原核生物蛋白质序列的预测方法（不同于 EXProt，一个蛋白质数据库）
- 真核生物蛋白质结构域预测方法[90]
- 革兰氏阳性细菌分泌性蛋白预测方法[91]
- 革兰氏阴性细菌 III 型分泌系统预测方法[52]

[a] 这些蛋白质在寄主与病原互作中起重要作用，它们可能是药物作用位点或候选疫苗。值得注意的是，目前的工具只能预测一部分分泌性蛋白或细胞表面蛋白。例如，目前还无法预测 IV 型分泌信号序列。可用于预测一般亚细胞定位（见文中叙述）的常用方法，有的基于 BLAST 等序列相似性分析，有的基于 InterPro 等模体和结构域分析。DAS 为 dense alignment surface 的缩写。

蛋白质定位分析有助于对微生物致病作用、治疗和预防靶位点的研究，在这方面有许多成功的例子（见框文 1），通过对脑膜炎球菌和导致牙周疾病的病原进行生物信息学分析，找出了细胞表面蛋白，从中发现了一些疫苗候选蛋白[36,50,51]。Pizza 等[51]的研究结果，首次说明基因组信息学在扩大和加速疫苗开发中的潜能。在他们的研究中，首先预测脑膜炎奈瑟氏球菌基因组序列中的表面蛋白，而后对这组蛋白进行大规模疫苗筛选，从中成功地得到 2 个候选疫苗。在这个过程中，他们采用了多种生物信息学方法（如 PHI-BLAST，FATSTA，MOTIFS，FIND-PATTERS，PSORT，ProDom，Pfam 和 Blocks)，去预测表面蛋白的典型特征，如跨膜结构域、信号肽、已知表面蛋白同源物、

脂蛋白特征、外膜蛋白锚定模体、寄主细胞结合结构域（如 RGD 序列：精胺酸-甘胺酸-天冬胺酸)[51]。许多研究组通常采用这些分析方法筛选疫苗，这也反映了目前还缺少鉴定疫苗的综合分析方法。

框文 1

以肺炎链球菌为例，说明生物信息学和实验室研究相结合，可鉴定新毒力因子和微生物基因组中抗感染位点。（更多的有关肺炎链球菌的例子，见 Di Guilmi 和 Dessen 的文献综述[92]。本章正文列举了其他病原的研究。）

肺炎链球菌是导致急性呼吸道感染和中耳炎的元凶，全球范围内每年可导致一百万人死亡。耐药性肺炎链球菌的出现和有效疫苗的缺乏要求去寻找新的有效抗生素和开发疫苗。该病原全基因组序列测定的完成，使得去研究成百上千个基因在感染过程中所发挥的作用，找出其生存所必需的基因，这些基因可能成为新的治疗靶位点。在肺炎链球菌菌株 TIGR4/N4 的基因组序列测定过程中，以及序列测定完成后，Wizemann[93]和 Tettelin[94]等计算、预测并分析了其中的可读框，从 2000 多个基因中寻找细胞表面蛋白或毒力因子。Wizemann 等的研究主要是寻找与转运蛋白、细胞壁蛋白、胆碱结合蛋白、整合蛋白等有关的蛋白质，以及与其他细菌细胞表面毒力因子具有相似性的基因。他们共找出 130 个可能的基因或基因片段，其中 108 个基因产物得到表达并纯化以作进一步研究；4 种基因产物，在弥散性肺炎链球菌小鼠感染模型中有保护作用，且在人体寄主中也具有免疫原性。还发现其中一种候选疫苗具有组氨酸三联体重复模体[95]，于是又对该菌全基因组序列进行检索，以寻找具有该特征的基因，结果又发现三个具有该特征的基因，其中两种基因产物为保护性免疫原。换言之，该策略首先对全基因组进行生物信息学分析，寻找可能的基因，表达和测试这些基因的产物，然后，根据得到的候选疫苗所具有的特征，再对该基因组进行分析，寻找可能的疫苗，最后对这些新发现的候选疫苗进行试验分析。

以上研究主要针对疫苗的筛选，筛选出的部分候选疫苗与细菌的毒力有关。而 Gosink 等[96]的研究是寻找肺炎链球菌的新毒力因子，根据 *cbpA* 基因羧基端胆碱结合区域所具有的特征，对肺炎链球菌全基因组进行分析，结果发现 6 个属于 *cbpA* 家族的基因。将这些基因逐个敲除之后，检测该菌对真核细胞的黏附性、在大鼠鼻咽的定殖能力和导致脓毒的能力，结果表明绝大多数与毒力有关，其中新发现的毒力因子 CbpG 与蛋白酶具有结构相似性，可能是一个很好的药物靶位点。

通过生物信息学分析，发现一些与亚细胞定位有关的新特征，这有助于鉴定微生物致病作用中起重要作用的蛋白质。例如 III 型分泌系统效应蛋白，在很多细菌病原毒力方面起重要作用，然而，通过常规生物信息学方法很难鉴定这些蛋白，原因是这些蛋白之间没有序列相似性。Guttman 等[52]分析了植物病原菌丁香假单胞菌中 13 种已知效应蛋白，结果发现这些效应蛋白的氨基端序列具有氨基酸组成相似性，因此开发出一种氨基酸组成分析方法，用该方法发现了 15 种可能的效应蛋白，后来的研究证明其中 2 种蛋白的分泌有赖于 III 型分泌系统。

Bannantine 等[53]发现，内含膜（inclusion membrane）上的衣原体蛋白有一种特殊的次级结构模体，根据这种结构特征，又新发现位于内含膜上的另外的蛋白质，并用抗血清方法验证了其定位。

预测表面/分泌性非蛋白化合物和分析代谢途径

除蛋白质外，微生物中很多化合物也位于细胞表面或分泌到细胞外，其中一些化合物就是毒力因子，如革兰氏阴性细菌中引起发热的脂多糖（内毒素）、多糖胶囊和非蛋白毒素（如硫化氢）。通常采用研究代谢途径的方法，并结合与相关代谢途径的基因进行同源分析，从而对这些化合物进行生物信息学研究。例如，为了鉴定全基因组中脂多糖的生物合成基因，通常采用 BLAST 方法，与已知脂多糖生物合成基因进行同源分析。然而，与特异代谢途径分析方法结合使用，可以弥补以上方法的不足。这种特异代谢途径分析方法，可以分析某代谢途径在某基因组中是否存在（见第 6 章）。

两种方法 MetaCyc 和 Pathway Tool 结合使用是其中一例[54,55]，Pathway Tools 是通过构建代谢途径、基因、蛋白质及其相关信息的数据库，对代谢和遗传网络进行分析、定义和展示。该软件包括四个主要子软件：PathoLogic、Pathway/Genome Navigator、Pathway/Genome Editor 和 Pathway Tools ontology。PathoLogic 是对已经注释的基因组创建一个新数据库，Pathway/Genome Navigator 是对已创建的数据库提供搜索、可视和网上服务，Pathway/Genome Editor 支持数据库更新，Pathway Tools ontology 是定义数据库结构。根据大肠杆菌菌株 K12 基因组创建的 EcoCyc 数据库，就是应用该方法的实例。

在创建 EcoCyc 数据库后，又创建了一个更普遍的代谢途径数据库 MetaCys（BioCyc）。2002 年首次公布时，该数据库包括 158 中生物中的 445 个代谢途径，1115 种酶[55]。这些代谢途径是综合多种文献的试验数据而确定，每种代谢途径的生物亦已标明，此外，微阵列分析数据也可置于这些代谢途径中，从而可对基因表达的变化进行代谢途径特异性分析，包括 KEGG 数据库[56]在内的资源，不仅对分析与毒力有关的代谢途径十分有用，而且有助于发现哪种代谢途径被阻断后，可导致生物毒力减弱或死亡，因此，这种分析对鉴定抗微生物药物的靶位点有帮助。

表面和分泌性化合物

鉴定相变基因

相变基因编码那些使细菌表型发生快速改变的蛋白质。这些基因十分有用，因为它们通常编码表面蛋白、分泌性蛋白或者合成表面化合物蛋白，这些表面化合物（如脂多糖）对致病性起关键作用。表型改变是通过开启或关闭某种或某类蛋白的表达而发生，这样，细菌可以不停地改变蛋白表达状态（蛋白表达开启或关闭，或表达同一基因家族中的不同亚种）。一般认为，这有利于细菌逃避寄主的免疫系统并快速适应新环境（如感染不同组织）。这种相变中的一种精细调控机制是同聚核苷酸重复（单核苷酸重复）和短重复序列的改变，通常位于编码基因或调控序列内。细菌通过同源短重复序列间的滑链错配机制发生遗传变异，这种变异可以影响基因的转录和蛋白质翻译，例如，改变位于基因编码序列起始点的同聚核苷酸重复数，可导致可读框的移码，使蛋白质的翻译提前终止，这种基因可读框的不断变化，控制该基因产物。请参阅 Moxon 等的综

述[57]，以便了解更多有关相变的例子。

全基因组序列测定后，可先鉴定基因组中的同聚核苷酸重复或短重复序列，再确定这些序列附近的基因，便可发现相变基因。Hood 等[2]首次采用该方法，对流感嗜血菌菌株 Rd 的基因组进行分析，寻找同聚核苷酸重复、二核苷酸重复、三核苷酸和四核苷酸重复的所有可能的组合。他们用 FINDPATTERN 软件[58]，搜寻所有同聚核苷酸重复和二核苷酸重复，而用 BLAST 搜寻三核苷酸和四核苷酸重复，然后，寻找这些短重复序列附近的基因。后来，Saunders 等[59]开发出更简捷的分析方法，该方法以 ACeDB 基因组分析和可视系统为基础，展示这些短重复序列，在基因组序列中所处的位置以及基因组的其他注解。采用该方法，在幽门螺杆菌菌株 26695 基因组序列中，发现了 10 个新相变基因[59]，在脑膜炎奈瑟氏球菌菌株 MC58 中发现了 52 个候选相变基因[60]。另外，还有其他短重复序列分析软件，如 EMBOSS（European Molecular Biology Open Software Suite）软件包中的 fuzznuc[61]。

这种通过短重复序列寻找相变基因的方法有很多优点，它可以找出真正的新毒力基因，因为该方法不是根据与已知毒力基因的同源性进行基因搜寻的。当然，该方法找出的基因也可能与毒力无关，而是发挥其他适应性优势。随着某个种或属内更多生物全基因组序列测定后，这种通过短重复序列鉴定相变基因的方法会更精细，因为那时就可以直接比较种间短重复序列的重复次数。某区域短重复序列重复次数的变异，能进一步说明该区域在相变过程中起作用，Snyder 等[62]根据这点对奈瑟氏球菌三个种的基因组进行综合生物信息学分析和短重复序列比较，结果发现了 100 多个可能的相变基因，另外还有一些假定基因，进一步研究发现，这些假定基因中有一个与 DNA 吸收感受态有关[63]。

鉴定抗原序列

免疫生物信息学（immunoinformatics）综合应用了信息学和免疫系统分子的模拟技术，它是快速发展的一个研究领域，本文只是提到而已，有关这方面的资源越来越多（如http://www.imtech.res.in/raghava/ctlpred/link.html 或http://www.jenner.ac.uk/bioinfo03/），而与免疫学有关的公开数据库成百上千[2,64]。免疫生物信息学最常用的分析技术之一是鉴定抗原序列（能在寄主体内引起免疫性反应的微生物蛋白序列），这些抗原序列具有作为疫苗的潜力，有助于揭示病原与寄主间的相互作用。

在进行抗原序列计算分析前，一般是先确定细胞表面蛋白或蛋白的表面暴露序列（如用表 1 所列举的亚细胞定位分析工具进行分析），再对其进行疏水性分析，以确定整个蛋白哪些区域最有可能位于蛋白质表面。通过以上分析，可将分析重点集中在最有可能被寄主免疫系统识别的蛋白序列上，如果细胞表面蛋白序列预测错误，当然会影响以后的分析工作，因此，在预测细胞表面蛋白或蛋白的表面暴露序列时，应使用高准确性的预测工具。

一旦某特定蛋白序列确定后，就可以研究它与寄主主要组织相容性复合体（MHC）类别分子的结合能力。因为，MHC 类别分子可以结合侵入寄主体内的微生物多肽片段，它在细胞免疫反应中起重要作用。由于 MHC 类别分子只能结合某特异多肽序列，于是开发了可提供有关 MHC 类别分子结合特异多肽的信息分析方法和数据库。有关这种特

异性结合最全面的数据库，包括 FIMM、SYFPEITHI 和 JenPep，其中 JenPep 可提供定量测定特异性结合亲和力，IMGT 是一个总结性信息系统[61]，可提供许多资源；ANTIGENIC：EMBOSS 是免费分析软件[65]，最初它是 GCG 生物信息学分析工具中的一个子软件；有些软件（如 nHLAPred 和 ProPred1）可同时预测 MHC 所结合的多肽和该多肽内的蛋白体切割位点。蛋白体切割是微生物抗原序列进入寄主免疫系统后的重要一步，通过蛋白体切割位点分析和 MHC 特异性结合分析，可以发现潜在的 T 细胞表位。BCIPep 是有关目前所有已报道 B 细胞表位的总结性数据库，通过

http://www.imtech.res.in/raghava/ctlpred/link.html 或

http://bioinformatics.uams.edu/mirror/mirror_imm.html 网站，进入 nHLAPred、ProPred1、BCIPep 以及其他预测工具和数据库。

与感染性疾病有关的人的多态性研究

尽管本章主要讨论用微生物基因组的分析方法研究微生物致病性，但是，微生物致病作用包括两方面：病原菌和寄主。

19 世纪后期，柯赫的研究结果阐明了一些疾病具有传染性。在这之前，普遍认为，如麻风病和结核病之类的疾病是遗传性疾病，但是，人对有些传染性疾病的易感性是可遗传的，这已经得到确认。而且，传染性疾病可以影响寄主的进化，镰状细胞贫血患者不易感染疟疾就是一个典型例子，在有地方性疟疾的地区，镰状细胞贫血等位基因比较常见，这是由于疟疾选择压对当地人群进行选择的结果（参阅综述文献 [66]）。

人类几个单基因突变引起的紊乱，改变了人对很多传染病的敏感性。例如，CD40 配体发生突变的个体对机会性感染比较敏感。人白细胞抗原（HLA）位点进化特别快，可能是病原选择压造成的结果。HLA 位点的多态性与人对传染病（如麻风病和结核病）的敏感性有关。人们还发现几个非 HLA 基因也与人对传染性疾病的敏感性有关，如维生素 D 受体基因影响人对结核病的敏感性，α 肿瘤坏死因子影响人对疟疾的敏感性，细胞因子 CD4 影响人对 HIV 的敏感性[66,67]。

随着人类基因组序列测定的完成，正在对影响人对传染病敏感性的基因（以及它们所伴随的多态性）进行更全面分析，这不仅有助于理解寄主遗传学在微生物致病过程中的作用，而且还可帮助开发出合适的治疗方法和疫苗。

应用美国生物技术中心的 dbSNP 数据库（http://www.ncbi.nlm.nih.gov/SNP），可直接从人类基因组中挖掘出与疾病有关的单核苷酸多态性（SNP）。通过 Ensebml (www.ensemble.org)[68]和其他网站都可以进入该数据库，这样就能以图解方式，预览各种基因组注释中的单核苷酸多态性，极大地方便了分析工作。

通过测定受某疾病影响（患过或正患该疾病）和未受该疾病影响个体（曾处于该疾病的环境中，但未患该疾病）中的候选基因，可发现在 dbSNP 数据库找不到的新单核苷酸多态性。如果从序列测定的原始数据中寻找新单核苷酸多态性，通常采用 Phred、Phrap 和具有 PolyPhred 的 Consed 系统工具（http://www.phrap.org），对自动测序产生的层析谱进行分析[69]。Phred 是广泛使用的软件，可推断出层析谱代表的序列，而对每个碱基产生一个质量值（quality score）；Phrap 是序列拼接软件，它可以将相互之间

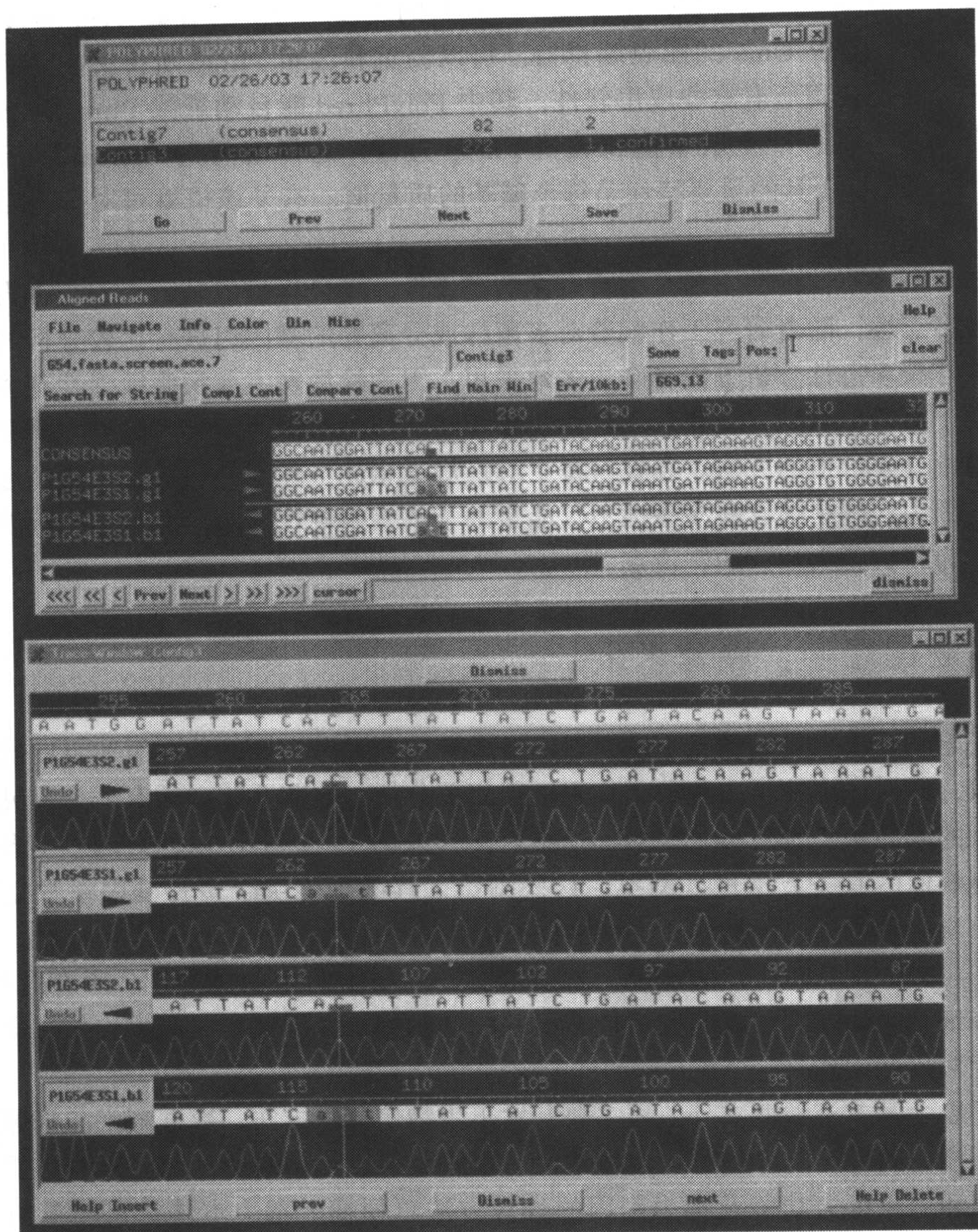


图 3　Phred、Phrap、Consed 和 Polyphred 分析软件，可帮助检测与人对某特定传染病的敏感性有关的单核苷酸多态性。本图采用 Consed 展示单核苷酸多态性。最上方 Polyphred 浏览窗口列举的重叠群，是通过运行 Phred（读取碱基）和 Phrap（拼接序列）而产生，其中一个序列来源于受疾病影响的个体，另一个来源于未受疾病影响的个体。将来自多个个体的核苷酸序列排列在一起，就能见到序列中的单核苷酸多态性，这种单核苷酸多态性可用 Polyphred 检测，用 Consed 将其展示。Polyphred 可将检测结果进行分级，1 表示单核苷酸多态性置信度高，6 代表置信度低。通过更改设置，可以改变检测单核苷酸多态性的敏感性和特异性。通过点击 Polyphred 浏览窗口中的重叠群，用户很容易用 Consed 浏览 Polyphred 检测出的单核苷酸多态性。本图列举了在重叠群 3 中检测的单核苷酸多态性，该单核苷酸多态性具有高置信度（置信度值为 1），并已得到确认（这里所说的确认是指 DNA 正向和反向链序列测定结果均表明单核苷酸多态性存在）。序列排列窗口展示了读取的核苷酸序列，大写字母为高质量读取的碱基，灰色背景的小写字母为低质量读取的碱基。请注意，真正单核苷酸多态性邻近的碱基，通常也注明为低质量读取的碱基，这是 Phred 产生质量值方法的本身原因，该方法考虑了侧翼碱基读取的质量。根据分级值的不同，检测出的单核苷酸多态性标上不同颜色。用户可点击感兴趣序列区域去浏览“追踪窗口”，该窗口会展示对应区域层析谱的原始数据。本图中以．g1 和．b1 为后缀的序列分别代表正向和反向序列，测定方向所获得的序列测定结果。S1 序列（P1G54E3S1）来源于单核苷酸多态性杂合子个体（在层析谱窗口中可以看到两个重叠的峰）。S2 序列（P1G54E3S2）来源于单核苷酸多态性纯合子个体。真正单核苷酸多态性（而非测序错误或者假象）的分级值，主要根据 Phred 产生的碱基质量值、两个重叠峰的面积比以及两个重叠峰的高度与纯合子峰理论上的高度比较结果而确定。

具有重叠序列的多个序列片段，拼接成一个更长的连续序列即重叠群（contig）；Consed和Autofinish是基于Unix的图表编辑器，可以浏览和编辑Phrap所拼接的序列。许多基因组中心广泛使用所有这些应用软件，其中PolyPhred能自动查找Phrap拼接序列中的单核苷酸多态性、产生质量值和编辑Phrap文件，这样就可用Consed展示多态性（图3），这套分析工具的重要特性是每个碱基的质量值，有了该值就可对分析结果进行定量评价。

单核苷酸多态性检测不只是分析寄主的多态性，还分析病原基因组中与毒力有关的单核苷酸多态性，随着研究工作朝着对多菌株或分离物的基因组进行比较分析的方向发展，单核苷酸多态性分析方法将会更多地采用。因为检测寄主与疾病敏感性有关的多态性，有助于更好地认识寄主和病原之间的复杂相互作用，因此，它可能成为研究微生物致病作用十分有价值的方法。寻找与寄主对疾病敏感性下降有关的单核苷酸多态性特别有意义，因为这些单核苷酸多态性不仅能帮助了解疾病抗性机制，而且它们所涉及的基因可能作为新抗感染治疗的靶位点。

病原生物信息学：展望未来

人们迫切需要大规模情报系统，以便对多种病原和非病原的基因组和蛋白质组进行自动比较，并对微生物毒力和致病作用进行相关分析。这些系统必须有一个更灵活、功能更强大的基因组浏览工具（如Gbrowse计划开发的工具[70]），以便对越来越多完成全序列测定病原的有关数据进行浏览。这些浏览工具必须能综合多种基因组数据（微阵列、蛋白质组和基因敲除数据），并能在更偏向细胞途径的水平浏览这些数据。为了对病原及其相关数据有更全面的展示，就要求有一个系统的生物学方法。所有这些问题，与分析多细胞生物基因组的研究者们所面临的问题相似，因此，希望生物信息学界能齐心协力应对这些问题。

除了以上基本资源要求外，还需要开发出生物信息学方法、数据词汇和病原微生物研究者所需的研究工具。如毒力本体论，有了它就可以创建真正的毒力基因数据库，并且使鉴定和分析毒力功能及毒力基因的计算方法更便捷。毒力本体论可以澄清毒力概念，能让生物信息学研究者更有效地研究毒力。

细胞表面定位有关序列的预测方法也有待进一步完善，特别是针对革兰氏阴性细菌和其他有外膜相关细菌的预测方法。由于定位在这些细菌外膜的蛋白彼此没有序列相似性，因此，对这些蛋白进行预测和鉴定十分困难（文献［71］中有这方面的例子）。随着预测方法由全蛋白水平朝逐个氨基酸水平方向发展，细胞表面蛋白和蛋白表面暴露序列预测方法的改进会越来越详细，这些改进的方法可更深入地分析哪个序列位于细胞表面。人们期望这样可以更准确地鉴定适合作为治疗靶位点或候选疫苗的蛋白，另外，还有待开发更多预测细胞代谢途径的预测方法，以便全面评价所有可能的细胞表面或分泌性化合物。

病原生物信息学是有待进一步挖掘的领域，包括开发可鉴定具有相似结构和功能蛋白的方法，如鉴定寄主-病原间相互作用蛋白的计算方法和模拟寄主-病原相互作用的方法（文献[72]中有这方面的例子），正在进一步完善鉴定定殖同一致病位点，病原间保

守序列的生物信息学系统（Fueyo J. L. 2003 年未发表资料），并且正在对与相变基因有关的短重复序列、与致病岛和其他水平转移序列（整合子）有关的特征进行更深入研究。

值得强调的是，目前病原生物信息学研究和发展受阻于一个明显的问题，尽管很多全基因组序列分析已经展开，但病原中哪些基因与毒力有关，这方面的综合分析还相对缺乏。因此，难以确定很多生物信息学方法在预测毒力基因方面的准确性，因为还无法知道哪些是真正阳性，哪些是真正阴性的结果。另一方面，“毒力”是个动态概念，它依赖于寄主的状态、病原的状态、感染条件和感染途径。然而，正如上所述，毒力本体论的发展是解决该问题的关键。

随着更多综合试验数据的出现和很多分析方法的发展，病原生物信息学将会更加量化，并在帮助进一步了解病原微生物和致病作用方面起重要作用。综合病原基因组学和生物信息学，可以发现病原致病作用的新线索、新治疗方法和预防靶位点。

（伍建宏，冯丽萍 译）

参考文献

1. World Health Organization. Report on Infectious Diseases: Removing Obstacles to Healthy Development. Atar, Switzerland: World Health Organization, 1999.
2. Hood DW, Deadman ME, Jennings MP, et al. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc Natl Acad Sci USA 1996; 93:11,121–11,125.
3. Nguyen L, Paulsen IT, Tchieu J, Hueck CJ, Saier MH Jr. Phylogenetic analyses of the constituents of type III protein secretion systems. J Mol Microbiol Biotechnol 2000; 2:125–144.
4. Holmes EC. Molecular epidemiology and evolution of emerging infectious diseases. Br Med Bull 1998; 54:533–543.
5. Levin BR, Lipsitch M, Bonhoeffer S. Population biology, evolution, and infectious disease: convergence and synthesis. Science 1999; 283:806–809.
6. Paine K, Flower DR. Bacterial bioinformatics: pathogenesis and the genome. J Mol Microbiol Biotechnol 2002; 4:357–365.
7. Zagursky RJ, Russell D. Bioinformatics: use in bacterial vaccine discovery. Biotechniques 2001; 31:636, 638, 640, passim.
8. Read TD, Gill SR, Tettelin H, Dougherty BA. Finding drug targets in microbial genomes. Drug Discov Today 2001; 6:887–892.
9. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403–410.
11. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988; 85:2444–2448.
12. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. Genomics 1991; 11:635–650.
13. Baxevanis AD, Ouellette BFF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. New York: Wiley, 2001.
14. Pertsemlidis A, Fondon JW 3rd. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol 2001; 2:REVIEWS2002.

15. Attwood TK, Bradley P, Flower DR, et al. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003; 31:400–402.
16. Apweiler R, Attwood TK, Bairoch A, et al. InterPro—an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 2000; 16:1145–1150.
17. Hacker J, Bender L, Ott M, et al. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. Microb Pathog 1990; 8:213–225.
18. Censini S, Lange C, Xiang Z, et al. cag, a pathogenicity island of *Helicobacter pylori*, encodes type I–specific and disease-associated virulence factors. Proc Natl Acad Sci USA 1996; 93: 14,648–14,653.
19. Ochman H, Soncini FC, Solomon F, Groisman EA. Identification of a pathogenicity island required for *Salmonella* survival in host cells. Proc Natl Acad Sci USA 1996; 93:7800–7804.
20. Blum G, Ott M, Lischewski A. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. Infect Immun 1994; 62:606–614.
21. Sullivan JT, Ronson CW. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc Natl Acad Sci USA 1998; 95:5145–5149.
22. Hentschel U, Hacker J. Pathogenicity islands: the tip of the iceberg. Microbes Infect 2001; 3: 545–548.
23. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H, et al. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 1997; 23: 1089–1097.
24. Inouye S, Sunshine MG, Six EW, Inouye M. Retronphage phi R73: an *E. coli* phage that contains a retroelement and integrates into a tRNA gene. Science 1991; 252:969–971.
25. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res 2002; 30:866–875.
26. Lio P, Vannucci M. Finding pathogenicity islands and gene transfer events in genome data. Bioinformatics 2000; 16:932–940.
27. Karlin S. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol 2001; 9:335–343.
28. Hsiao W, Wan I, Jones SJ, Brinkman FS. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 2003; 19:418–420.
29. Koski LB, Morton RA, Golding GB. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 2001; 18:404–412.
30. Sonnhammer ELL, Durbin R. A workbench for large scale sequence homology analysis. Comput Appl Biosci 1994; 10:301–307.
31. Florea L, Riemer C, Schwartz S, et al. Web-based visualization tools for bacterial genome alignments. Nucleic Acids Res 2000; 28:3486–3496.
32. Schwartz S, Kent WJ, Smit A, et al. Human–mouse alignments with BLASTZ. Genome Res 2003; 13:103–107.
33. Perrin A, Bonacorsi S, Carbonnelle E, et al. Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. Infect Immun 2002; 70:7063–7072.
34. Read TD, Salzberg SL, Pop M, et al. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 2002; 296:2028–2033.
35. Paulsen IT, Seshadri R, Nelson KE, et al. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci USA 2002; 99: 13,148–13,153.

36. Tettelin H, Masignani V, Cieslewicz MJ, et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci USA 2002; 99:12391–12396.
37. Fleischmann RD, Alland D, Eisen JA, et al. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 2002; 184:5479–5490.
38. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 2001; 409:529–533.
39. Fueyo JL. In Silico Discovery of Antimicrobial Targets. Ph.D. thesis, Philadelphia: University of Pennsylvania, 2002.
40. Bruccoleri RE, Dougherty TJ, Davison DB. Concordance analysis of microbial genomes. Nucleic Acids Res 1998; 26:4482–4486.
41. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988; 85:2444–2448.
42. Thanassi JA, Hartman-Neumann SL, Dougherty TJ, Dougherty BA, Pucci MJ. Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. Nucleic Acids Res 2002; 30:3152–3162.
43. Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc Natl Acad Sci USA 2002; 99:1556–1561.
44. Galperin MY, Koonin EV. Searching for drug targets in microbial genomes. Curr Opin Biotechnol 1999; 10:571–578.
45. Huynen M, Dandekar T, Bork P. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 1998; 426:1–5.
46. Tatusov RL, Natale DA, Garkavtsev IV, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001; 29:22–28.
47. Emanuelsson O. Predicting protein subcellular localisation from amino acid sequence information. Brief Bioinform 2002; 3:361–376.
48. Gardy JL, Spencer C, Wang K, et al. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 2003; 31:3613–3617.
49. Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci 2002; 11:2836–2847.
50. Ross BC, Czajkowski L, Hocking D, et al. Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. Vaccine 2001; 19:4135–4142.
51. Pizza M, Scarlato V, Masignani V, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000; 287:1816–1820.
52. Guttman DS, Vinatzer BA, Sarkar SF, Ranall MV, Kettler G, Greenberg JT. A functional screen for the type III (Hrp) secretome of the plant pathogen *Pseudomonas syringae*. Science 2002; 295:1722–1726.
53. Bannantine JP, Griffiths RS, Viratyosin W, Brown WJ, Rockey DD. A secondary structure motif predictive of protein localization to the chlamydial inclusion membrane. Cell Microbiol 2000; 2:35–47.
54. Karp PD, Paley S, Romero P. The Pathway Tools software. Bioinformatics 2002; 18(Suppl 1): S225–S232.
55. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The MetaCyc database. Nucleic Acids Res 2002; 30:59–61.
56. Kanehisa M. The KEGG database. Novartis Found Symp 2002; 247:91–101; discussion 101–103, 119–128, 244–252.
57. Moxon ER, Rainey PB, Nowak MA, Lenski RE. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol 1994; 4:24–33.

58. Devereux J, Haeberli P, Smithies O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 1984; 12:387–395.
59. Saunders NJ, Peden JF, Hood DW, Moxon ER. Simple sequence repeats in the *Helicobacter pylori* genome. Mol Microbiol 1998; 27:1091–1098.
60. Saunders NJ, Jeffries AC, Peden JF, et al. Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. Mol Microbiol 2000; 37:207–215.
61. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000; 16:276–277.
62. Snyder LA, Butcher SA, Saunders NJ. Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria* spp. Microbiology 2001; 147:2321–2332.
63. Snyder LA, Saunders NJ, Shafer WM. A putatively phase variable gene (dca) required for natural competence in *Neisseria gonorrhoeae* but not *Neisseria meningitidis* is located within the division cell wall (dcw) gene cluster. J Bacteriol 2001; 183:1233–1241.
64. Brusic V, Zeleznikow J, Petrovsky N. Molecular immunology databases and data repositories. J Immunol Methods 2000; 238:17–28.
65. Lefranc MP. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol 2002; 26:697–705.
66. Cooke GS, Hill AV. Genetics of susceptibility to human infectious disease. Nat Rev Genet 2001; 2:967–977.
67. Foster CB, Chanock SJ. Mining variations in genes of innate and phagocytic immunity: current status and future prospects. Curr Opin Hematol 2000; 7:9–15.
68. Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucleic Acids Res 2002; 30:38–41.
69. Nickerson DA, Tobe VO, Taylor SL. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 1997; 25:2745–2751.
70. Stein LD, Mungall C, Shu S, et al. The generic genome browser: a building block for a model organism system database. Genome Res 2002; 12:1599–1610.
71. Brinkman FS, Bains M, Hancock RE. The amino terminus of *Pseudomonas aeruginosa* outer membrane protein OprF forms channels in lipid bilayer membranes: correlation with a three-dimensional model. J Bacteriol 2000; 182:5251–5255.
72. Stebbins CE, Galan JE. Structural mimicry in bacterial virulence. Nature 2001; 412:701–705.
73. Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 1998; 283:489–506.
74. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001; 305: 567–580.
75. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 1997; 10:673–676.

75a. Hoffman K, Stoffel W. TMbase—a database of membrane spinning proteins segments. Biol Chem Hoppe-Seyler 1993; 374:166. http://www.ch.embnet.org/software/TMPRD_form.html.

76. Janulczyk R, Rasmussen M. Improved pattern for genome-based screening identifies novel cell wall-attached proteins in Gram-positive bacteria. Infect Immun 2001; 69:4019–4026.
77. Diederichs K, Freigang J, Umhau S, Zeth K, Breed J. Prediction by a neural network of outer membrane beta-strand protein topology. Protein Sci 1998; 7:2413–2420.
78. Jacoboni I, Martelli PL, Fariselli P, De Pinto V, Casadio R. Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. Protein Sci

2001; 10:779–787.
79. Martelli PL, Fariselli P, Krogh A, Casadio R. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 2002; 18(Suppl 1):S46–S53.
80. Zhai Y, Saier MH Jr. The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. Protein Sci 2002; 11:2196–2207.
81. Schirmer T, Cowan SW. Prediction of membrane-spanning beta-strands and its application to maltoporin. Protein Sci 1993; 2:1361–1363.
82. Wimley WC. Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci 2002; 11:301–312.
83. Horton P, Nakai K. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. Proc Int Conf Intell Syst Mol Biol 1997; 5:147–152.
84. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 2002; 18:298–305.
85. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997; 8:581–599.
86. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001; 17:721–728.
87. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998; 26:2230–2236.
88. Nakai K, Kanehisa M. Expert system for predicting protein localization sites in Gram-negative bacteria. Proteins 1991; 11:95–110.
89. Saleh MT, Fillon M, Brennan PJ, Belisle JT. Identification of putative exported/secreted proteins in prokaryotic proteomes. Gene 2001; 269:195–204.
90. Mott R, Schultz J, Bork P, Ponting CP. Predicting protein cellular localization using a domain projection method. Genome Res 2002; 12:1168–1174.
91. Gomez M, Johnson S, Gennaro ML. Identification of secreted proteins of *Mycobacterium tuberculosis* by a bioinformatic approach. Infect Immun 2000; 68:2323–2327.
92. Di Guilmi AM, Dessen A. New approaches towards the identification of antibiotic and vaccine targets in *Streptococcus pneumoniae*. EMBO Rep 2002; 3:728–734.
93. Wizemann TM, Heinrichs JH, Adamou JE, et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect Immun 2001; 69:1593–1598.
94. Tettelin H, Nelson KE, Paulsen IT, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 2001; 293:498–506.
95. Adamou JE, Heinrichs JH, Erwin AL, et al. Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis. Infect Immun 2001; 69:949–958.
96. Gosink KK, Mann ER, Guglielmo C, Tuomanen EI, Masure HR. Role of novel choline binding proteins in virulence of *Streptococcus pneumoniae*. Infect Immun 2000; 68:5690–5695.

5

噬菌体生物信息学

Derrick E. Fouts

引言

我们正逐渐了解到溶原性噬菌体为它们的寄主提供必要的毒力和适应因子，影响细菌附着、定居、侵入、扩散、免疫应答抗性、外毒素产生、血清抗性及抗生素抗性。应用噬菌体编码降解细菌细胞壁和抑制细胞壁前体合成酶的最新研究，为更新噬菌体为基础的药物提供了可能性，此外，通过研究噬菌体基因组，可以对细菌基因组的可塑性和进化有更多的了解。

1959 年以来，已经对 5000 多种噬菌体进行了分类[1]，然而，完全测序且其序列能在基因库中公开得到的噬菌体不到其 3%（Entrez Genomes：Phages；表 1）。为了充分了解噬菌体基因组的多样性、基因移动或交换的全部潜能及其进化，需要更多的噬菌体基因组序列。与此同时，噬菌体研究者们能挖掘大量、很少探索和公众可利用的资

表 1　有用的网站连接

描述	网址
Artemis	http://www.sanger.ac.uk/Software/Artemis/
CLUSTAL W	http://www-igbmc.u-strasbg.fr/BioInfo/ClustalW/clustalw.html
Comprehensive Microbial Re-source(CMR)	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
Entrez Genomes: Phages	http://www.ncbi.nlm.nih.gov:80/genomes/static/phg.html
HMMER	http://hmmer.wustl.edu/
International Committee on Taxonomy of Viruses (ICTV)	http://www.ncbi.nlm.nih.gov/ICTVdb/
Manatee(Manual Annotation Tool, etc., etc.)	http://manatee.sourceforge.net/
Pfam	http://pfam.wustl.edu/
Phage Ecology	http://www.mansfield.ohio-state.edu/~sabedon
Phage genomics(international Bacteriophage Genomics Group)	Http://meds.queensu.ca/~ibgg/
Phage proteome	http://salmonella.utmem.edu/phage/tree/
Phage therapy	http://www.evergreen.edu/phage/
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
PSI-BLAST	http://www.ncbi.nlm.nih.gov/blast
REPuter	http://www.genomes.de/
TIGRFAMs	http://www.tigr.org/tigr-scripts/CMR2/find_hmm.spl?db=CMR
WU-BLAST(Washington Univer-sity BLAST Archives)	http://blast.wustl.edu

源——全细菌基因组。本章不打算反复讲述已经阐述清晰的噬菌体分类学[2,3]、进化[4-8]或比较噬菌体基因组学[9~18]，而重点对完整或不完整细菌基因组原噬菌体的鉴定和分析的生物信息学进行阐述。

背景

噬菌体是侵染细菌的病毒，通常采用它们在易感的或寄主菌苔上产生的噬菌斑进行鉴定。噬菌体一词是希腊语“食”的意思，因此，噬菌体就是食细菌者。出人意料的是，噬菌体并非是第一个被鉴定的病毒，动物和植物病毒获得了这一殊荣，但记录描述的病毒中，噬菌体数量最大[1]。谁首先发现噬菌体是细菌病毒有争论，大约在 1915～1917 年，不是法裔加拿大细菌学家 Felix d'Herelle，就是英国病理学家 F. W. Twort[19,20]，也不知道谁是第一个鉴定了噬菌体，要么是感染引起痢疾的细菌（志贺氏菌），要么是引起蝗虫肠炎的细菌（一种产气肠杆菌的亚种）[19~23]，直到 20 世纪 40 年代早期，电子显微镜才第一次观察到噬菌体的结构[24~28]。

直至今日，细菌遗传学和分子生物学中的很多重要原则都是根据噬菌体研究结果进行阐述的，例如，Luria 和 Delbrück 应用著名的波动试验，证明了细菌在用 α 噬菌体（现在称 T1）作为选择压力前，能够随机获得突变[29]。T1 是 7 个感染大肠杆菌（*E. coli*）菌株 B 的裂解性噬菌体之一（T1-T7）。1944 年 Delbrück 和冷泉港实验室噬菌体研究组，将这些噬菌体放在一起作为 T 系统的一部分，以便集中和标准化对裂解性噬菌体进行研究[28]。证明脱氧核糖核酸是遗传物质的关键实验之一，就是采用噬菌体 T2 做的著名 Hershey-Chase 实验[30]。转导概念是 Zinder 和 Lederberg 通过噬菌体 PLT22（现在叫 P22）实验而建立的，转导即噬菌体将一种寄主的基因组 DNA 转移到另一种寄主中去的能力，是构建具有特定遗传背景细菌菌株的常见方法[31]。利用称为 rⅡ 的 T4 寄主范围突变体（快速裂解Ⅱ型突变株），Benzer 用缺失作图定义了一个基因的物理结构[32]。遗传密码的破译和确认是通过研究 T4 噬菌体而实现的[33,34]。在分子生物学中常规使用的许多酶和技术都源于噬菌体研究：限制性内切酶[35]、DNA 连接酶[36,37]、RNA 连接酶[38]、DNA 聚合酶[39]、多核苷酸激酶[40]、噬菌体展示[41]，攻击噬菌体（challenge phage）[42,43]、M13 包装单链 DNA 测序分析[44,45]和定点诱变[46,47]以及 λ 克隆载体[48]。

裂解性或烈性噬菌体是只经历一个生活周期（营养生长）的噬菌体。该生活周期包括：黏附寄主细胞、病毒核酸侵入、劫持寄主代谢机制、复制自己，并摧毁细胞释放成百上千新合成的病毒粒子。溶原性或温和噬菌体却能经历另外一个生活周期，即通过将 DNA 结合到寄主染色体上，作为原噬菌体而停留下来。原噬菌体一般处于转录静止的休眠阶段，只产生编码阻抑蛋白的信使核糖核酸（mRNA）和其他可对原噬菌体适应性起作用的一些基因。[7,17]。

对噬菌体研究了 50 多年才建立分类学标准[49]，要了解当今噬菌体分类学概况，可参阅文献［1］和［3］。有尾病毒目（*Caudovirales*）是具尾的噬菌体，该目包含大部分已知噬菌体，该类群的一个亚群是温和型类群，所有具尾噬菌体都含有双链线状 DNA，根据尾部形态学将其分为三个科：肌尾病毒科（*Myoviridae*），有可收缩的尾部，

像典型的注射器模样，如T2噬菌体；长尾病毒科（*Siphoviridae*），有长而不收缩的尾部，像λ噬菌体；短尾病毒科（*Podoviridae*）有短的尾部，如P22。这三个科中的每个科可再分成三种形态类型（如肌尾病毒科分成从A1到A3，长尾病毒科可分为B1到B3，短尾病毒科可分为C1到C3）。每个数字代表衣壳形状：1～3分别代表正面体、稍长和很长的头部。噬菌体还可根据其他特征来鉴定，包括核酸基因组类型[3]。

对那些能增加细菌适合度的基因，噬菌体是这些基因遗传流动性的载体[7,13]，特别是早在1927年，噬菌体在人类细菌致病机制中的作用就受到重视和研究[50,51]。实验证明噬菌体携带的基因，对细菌毒力方面（附着、侵入、寄主躲避和毒素产生）有许多作用[51]，毒素好像是噬菌体产生能观察到最广泛的毒力因子。例如，噬菌体携带有肉毒杆菌毒素[52]、白喉毒素[53,54]、霍乱毒素[55]、志贺氏菌毒素[56,57]和毒性休克综合性毒素[58]。有人提出轻型链球菌（*Streptococcus mitis*）噬菌体SM1编码的PblA和PblB蛋白质，有助于噬菌体附着在人类血小板上，可能导致心内膜炎[59]。噬菌体通过改变细菌O抗原结构[60,61]或提供一些能中和氧化性裂解的酶——这种氧化裂解酶可破坏入侵的细菌[62]，使其细菌寄主能侵入人类免疫系统，噬菌体也有一些编码酶，尤其是对细菌毒力起直接作用的酶[63~65]。

如果已经充分论述的噬菌体，在细菌致病病理中的重要作用，作为扩大噬菌体研究的理由还不充足，那么，采用噬菌体治疗人类疾病的想法将更受欢迎[66~68]。在抗生素发现前，传统的噬菌体疗法曾经是一种可接受处理某些细菌感染的方法。例如，从感染急性痢疾病人的大便中可分离出噬菌体，将这些噬菌体培养到很高的滴度，并让病人吞食，在某些情况下，这种治疗疾病的方法十分有效[69,70]。最近的噬菌体疗法治愈了老鼠由抗万古霉素的粪肠球菌（*Enterococcus faeciium*）引起的菌血症[71]。噬菌体治疗失败的案例多是因为针对靶细菌选择的噬菌体不适宜，或是细菌对某种噬菌体产生了抗性，或是因为噬菌体携带有导致细菌致病的毒性因子。

前苏联医学专业人员专门针对感兴趣的细菌配制了含不同噬菌体、无细菌污染的制剂[72]。Fischeti小组的研究刊在《自然》杂志的封面。据报道，从感染炭疽芽孢杆菌（*Bacillus anthracis*）的γ噬菌体中，提取纯化的细胞溶素蛋白（PlyG）能用于检测和破坏炭疽芽孢杆菌萌发的细胞[73]。这种噬菌体疗法的新转变证明是有用的，因为与用整个噬菌体裂解物相比，该方法使细菌产生抗性的频率低很多。溶素蛋白倾向于识别细胞壁组分，这种组分的变异容易对细菌产生负面影响，而典型噬菌体受体分子的变异就简单得多，只要增加或减少受体分子的糖残基或改变外膜蛋白就可以了。

噬菌体产品另一潜在医药用途是基因治疗，最近已证明，噬菌体的位点特异性整合酶在人细胞中起作用，这使得用工程整合酶将引起疾病等位基因的正常拷贝转移到人类基因组或其他哺乳动物基因组中的安全区域在理论上成为可能[74~78]。

基因组学时代给很多生物体的编码能力和基因组的可塑性带来了前所未有的知识财富，噬菌体对这个时代的贡献经常因没得到应有认识而被忽略。基因组学并不是从1995年流感嗜血菌（*Haemophilus influeirzae*）的测序才开始的[79]，因为ΦX174和λ噬菌体基因组分别于1977和1983年测序[80~83]。噬菌体基因组测序导致了细菌基因组的测序，具讽刺意味的是，通过整合原噬菌体形式，细菌基因组正作为获得噬菌体基因组资料的一种途径。

从细菌基因组中发现原噬菌体

阻碍准确鉴定细菌基因组中原噬菌体区段的最大挑战之一，是对噬菌体基因的适当注释。例如，为了注释具有某一特殊功能的基因，必须满足基因组研究所（TIGR）的以下标准：必须有 1×10^{-5}或更小的 BLAST（Basic Local Alingment Search Tool）（基本局部联配搜索工具）E 值，必须与用实验方法已研究清楚的蛋白质有至少 35%氨基酸一致性，或者未知蛋白的得分要大于隐式马可夫模型（Hidden Markov Model，HMM）所确定的某一可信度阈值，因为这个阈值是根据参考蛋白的实验研究而建立的，即使有很低的 BLAST E 值和可接受的一致性百分比。如果这种蛋白质的功能特性还未在实验室内确定，则在推定原噬菌体区域中的基因可能被注释为“保守的假定蛋白”——正像从细菌基因组计划中寻找原噬菌体的匹配一样。

还有些情况是，基因特性研究得很清楚，但是参考文献发表很早或发表在在线文献资源无法查找的杂志上，致使在线文献资源中找不到参考文献。通常，在原噬菌体区域中，一种基因编码的假设蛋白质并没有明显的 BlAST 或马可夫模型匹配，正是这种欠缺而被标注为假定蛋白，这可能由于公众数据库中缺乏特征性噬菌体蛋白的代表和噬菌体基因库存在巨大的遗传多样性。尤其麻烦的是，在 20～30kbp 区段内，有大量可读框被注释为有噬菌体或噬菌体样功能，但是，该区段并没有标记为连续的原噬菌体区段，换句话说，局部 ORF 的遗传结构或相对位置被忽视了，使得整合原噬菌体内的 ORF 无法识别。

在详细考虑原噬菌体区域鉴定前，先定义一下我对原噬菌体区域含义的理解。首先，必须提到，在通过诱导分离出感染噬菌体粒子前，所有原噬菌体区域都是假定的原噬菌体区域，原噬菌体区域是一簇基因或一段基因，这些基因可能编码具有噬菌体样功能的蛋白，这些基因与功能未知或已知噬菌体编码基因交错排列。这些区域可能与噬菌体的复制、形态发生、组装、免疫或寄主裂解有关，也可能无关。关于真正的原噬菌体、缺陷的或隐蔽的原噬菌体，甚至更令人迷惑的噬菌体样的细菌素都有许多不同的解释[84~86]。

真正的原噬菌体是可以诱导的，而缺陷型原噬菌体不可诱导，它是功能性温和噬菌体的残骸。我们所说的隐型原噬菌体（cryptic prophage），是历史遗留缺陷的原噬菌体，它们不能提供针对超感染噬菌体的免疫性[87]。分类中另一要考虑的是通过非法重组（illegitimate recombination）方式，整合到细菌寄主基因组中的烈性噬菌体基因[88]。在寄主和噬菌体之间，进行着战争与和平的连续循环，寄主想保留原噬菌体编码特定功能的特性，但又不想让原噬菌体诱导引起致死，因而，谁先杀死谁在时间上进行竞赛[89]。

可能存在一种似乎有功能但某个关键基因发生了突变的原噬菌体。例如，在化脓链球菌（*Streptococcus pyogenes*）370.2 原噬菌体区段中[90]，门户蛋白（portal protein）的突变可能使原噬菌体失去活性。后期经插入/缺失和重组引起的突变，产生了一个退化原噬菌体区段，该缺陷型原噬菌体区段很容易被识别。预测卫星原噬菌体（satellite prophage）要有一点技巧，卫星原噬菌体是没有能力进行裂解的原噬菌体，除非有完全功能、定居在寄主内的原噬菌体或外来噬菌体，通过反式作用提供功能性组分，卫星噬

菌体才具有裂解能力[9]。

有功能的最小双链DNA温和有尾噬菌体之一（枯草芽孢杆菌Φ29-样噬菌体B-103）的大小约为18kbp[91]，而试验证明，乳酸乳球菌（*Lactococcus lactis*）的卫星噬菌体为13～15kbp[9]，所以，一直用10～18kbp作为潜在卫星噬菌体大小的范围，这对较大原噬菌体可能适用，也可能不适用。比18kbp大的区段均可认为是假设的原噬菌体，直至有其他观察研究怀疑这个界定。为了这个目的，我没有包含任何小于10kbp的噬菌体基因类群，这并不认为噬菌体样整合酶就代表缺陷原噬菌体区段，因为它可能是由寄主衍生，或来自其他可移动的遗传组分[92]。

最后，其他原噬菌体区段就是噬菌体样的细菌素，铜绿假单胞菌R-型或F-型的绿脓菌素就是一个例子[93]，这些区段容易被认为是缺陷原噬菌体，因为它们确实有很多类似已知噬菌体基因的基因（R型绿脓菌素中的P2和F型绿脓菌素中的λ噬菌体)。噬菌体样的细菌素不同，它们含有编码尾丝和附属物的基本基因、调节蛋白的基因和裂解基因，但没有编码外壳蛋白基因。这些区段失去了所有与尾丝形成无关的基因，被寄主用来破坏其他受体细胞的细胞膜而杀死类似种类的细菌[94]。这些细菌素有很多不同名称，如单核细胞利斯特氏菌（*Listeria monocytogenes*）产的利氏菌素，鼠疫耶尔森氏菌（*Yersinia pestis*）产的鼠疫菌素，蜡状芽孢杆菌（*Bacillus cereus*）产的蜡样菌素，大肠杆菌产的大肠菌素和葡萄球菌（*Staphylococcus*）产的葡萄球菌素[95]。如果有一段噬菌体样基因簇，它不包含衣壳体基因，但包含很多拟尾基因，该区段可能是一种噬菌体样细菌素，像噬菌体一样，这些细菌素对病原细菌菌株的分型或鉴定有用[96,97]。

由于鉴定原噬菌体有困难，就必须对每个区段进行人工检测，这需要用软件工具观察一些与已知噬菌体基因，有明显BLAST匹配的基因附近的遗传组成，这些软件工具有图形化观察可读框的软件，如Artemis（表1）或用TIGR-Manatee注释工具包中的区段显示选择功能（表1)。原噬菌体区段的界限要通过浏览感兴趣基因的5′到3′端，直到发现像糖酵解酶或RNA核糖体蛋白此类“管家”基因。可以用如REPfind工具（一种对公众有用的REPuter软件包)[98]来研究原噬菌体区段中假定边界两端扩展区(～1kbp)的序列重复性，以寻找DNA整合位点之间由单交换整合而产生的正向重复片段[99]，其中一个拷贝是噬菌体衍生的，另一个是寄主衍生的。

这种决定插入或整合位点的方法，对Mu样噬菌体（Mu-like phage）不起作用，因为这些噬菌体通过位点转座机制整合到基因组上，导致产生一个4～5bp的靶位点重复[100]。相反，Mu样噬菌体区段的边界，可通过功能保守基因的定位来粗略鉴定，噬菌体区段末端有一个cI样的阻抑物（Mu中*c*功能的类似物)，附近有两个类似转座酶A和B亚单位的可读框，这一区段的另一端，可以根据与特异性腺嘌呤甲基化酶（与Mu中*mon*功能相当）具有相似可读框的存在而确定[100]。脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）血清B类群菌株MC58的MuMen B，似乎不同于以前公布的4种Mu样噬菌体，该原噬菌体是通过破坏ABC转运蛋白酶NMB1077/NMB1122而偶然发现[101]。这种原噬菌体产生一个7bp不完整靶位点重复，并含有cI样阻抑物（蛋白)、转座酶A和B亚单位，但没有*mom*等同物，所以，这种噬菌体边界只能通过打断寄主基因来确定。

如果没有发现正向重复，那么可以采用其他措施。如果有多个亲缘细菌的完整基因

组序列，可以在假设原噬菌体区段附近寻找打断的保守基因，正如对炭疽芽孢杆菌（*B. anthracis*）的假设原噬菌体区 3 所做的那样（图 1）。在这个例子中，没有找到潜在 attL/R 位点这样的正向重复，因此，位于手工鉴定的原噬菌体区段两侧的基因，可在已发表的下列细菌基因组中找出来比较，它们是枯草芽孢杆菌（*B. subtilis*）、耐盐芽孢杆菌（*B. halodurans*）、单核细胞利斯特氏菌（*L. monocytogenes*）和金黄色葡萄球菌（*Staphylococcus aureus*）。结果发现，在 *ylbM* 和 *ylbN* 基因之间发生过一次或多次的原噬菌体整合，因为 *ylbL*、*ylbM*、*ylbN* 和 *rpmF* 的基因顺序（枯草芽孢杆菌的基因名称）在炭疽芽孢杆菌中被打乱（图 1）。

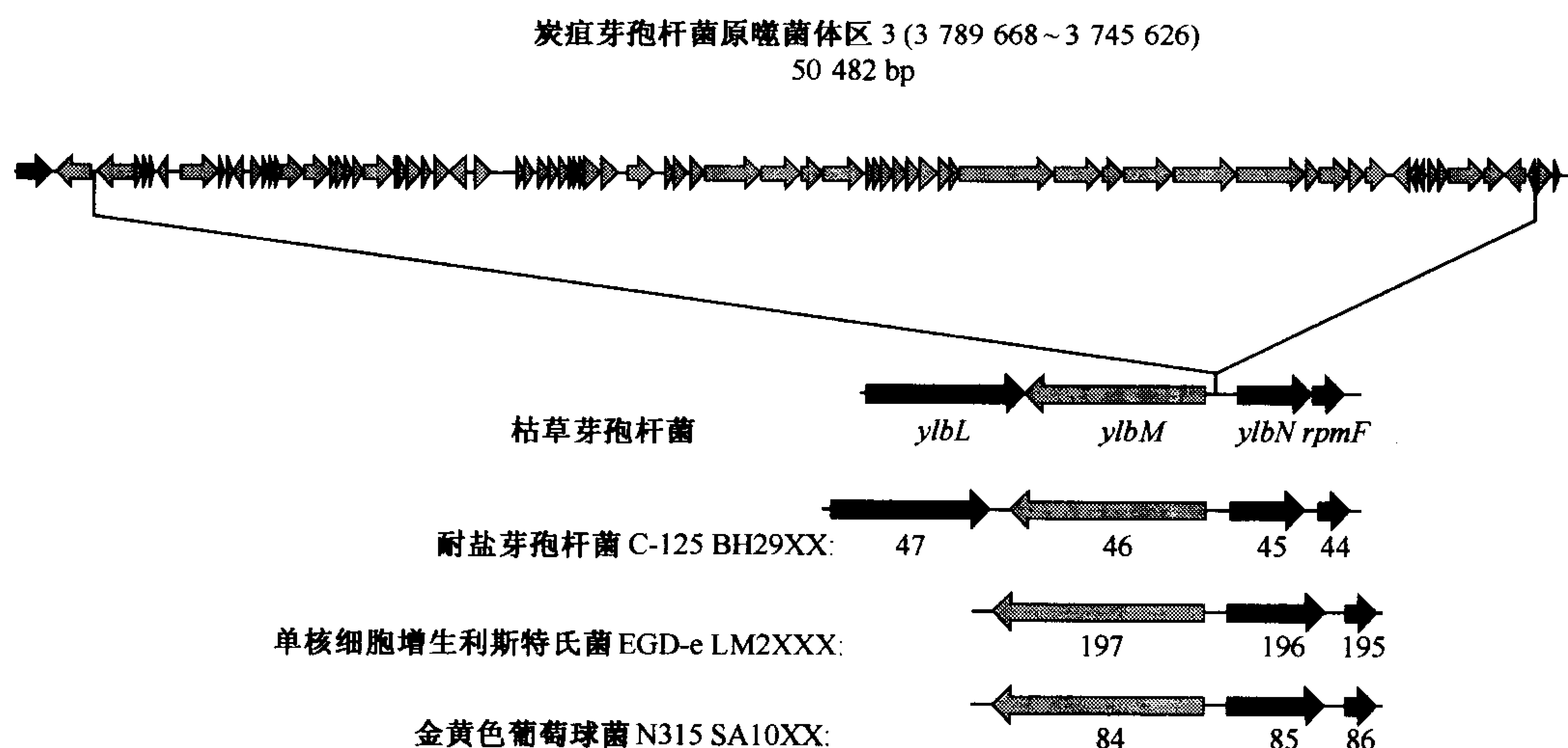


图 1 细菌基因组的原噬菌体边界，可以通过破坏已鉴定近缘微生物的保守基因顺序来确定。枯草芽孢杆菌和耐盐芽孢杆菌（*Bacillus haloduran*）基因组之间，*ylbL*、*ylbM*、*ylbN* 和 *rpmF* 顺序显然是保守的。更大范围看，在 G^+ 细菌枯草芽孢杆菌、耐盐芽孢杆菌、单核细胞增生利斯特氏菌和金黄色葡萄球菌中，*ylbM*、*ylbN* 和 *ylbF* 的顺序显然是保守的。相关可读框名称在每个可读框下标明。

在原噬菌体的 3′端，出现一个假定管家基因 *spoⅢE*（BA4067）的解释是：从炭疽芽孢杆菌或蜡状芽孢杆菌的亲缘菌株经转导而来，蜡状芽孢杆菌是与炭疽芽孢杆菌亲缘关系密切的菌株，支持该理论的证据是在炭疽芽孢杆菌中有 2 个拷贝的 *spoⅢE*，而在枯草芽孢杆菌和耐盐芽孢杆菌中只有一个拷贝。由于产气荚膜梭菌的噬菌体 ϕ3626 编码了一种类似 SpoⅢD 的蛋白质[102]，可以推测，这些假定噬菌体编码的转录调节蛋白，操纵着它们寄主内生芽孢的发育。然而，这些蛋白也可能是一些未知或特征不明确的噬菌体功能调节蛋白，该区段是否可诱导，这些假定 sigma 因子是否在炭疽芽孢杆菌的适应性或毒力上起作用，还需要实验研究进一步确定。

对于含有可识别整合酶基因的原噬菌体区段，靶位点很可能靠近整合酶基因[99]，如果整合酶基因附近碰巧是一个转运 RNA（tRNA）基因，那么，可能就是靶位点（attB）。因为，酪氨酸重组酶家族的很多整合酶都有 tRNA 基因的倾向性[99]。这种假定的靶位点，可用来寻找在假设噬菌体区段另一边的另一个类似拷贝[103]，如果假定靶位点是一个 tRNA 基因，像 Lowe 和 Eddy 的 tRNAscan-SE 这样的程序，就可找到这个配

对拷贝[104]。靶位点不总是 tRNA 或转运信使 tmRNA，它们可能像 6 磷酸葡萄糖异构酶[105]和鸟苷-5′-单磷酸（GMP）合成酶这样的保守管家基因[102]。或许噬菌体整合到保守靶位点上有利，这可能使噬菌体通过确保有个“登陆垫”而增加寄主范围，因为无论哪种细菌都可能碰到适宜的细胞表面受体。由于噬菌体一般不可能使寄主生活所必需的基因失活，因此，很难发现导致靶基因移码的整合原噬菌体，不能依赖鸟嘌呤和胞嘧啶含量（G+C%）或非典型核苷酸组分来寻找原噬菌体，因为有尾噬菌体的 G+C 含量与它们寄主的 G+C 含量相当[3]，例外的是，恶臭假单胞菌的假定原噬菌体正好位于基因组的非典型区段[106]。

手工测定原噬菌体区段很费时费力，特别是当分析规模达到当前保存在 TIGR 综合微生物资源库（TIGR's Comprehensive Microbial Resource，CMR）的 90 多种细菌基因组时[107]。为了解决这个问题，用 Phage-phinder（噬菌体发现器）来发现细菌基因组中的原噬菌体区域，并以多种格式文件输出结果。Phage-phinder 是一种由 Perl 语言（Practical Extraction and Report Language，实用提取和报告语言）编写的软件。先用 BLASTP 2.0MP-WashU[108]程序在细菌基因组中寻找所有已知的噬菌体序列，然后用定位分隔（tab-delimited）形式把搜索结果输入 Phage-phinder 程序，如果某基因与噬菌体 BLAST 数据库中的条目有显著匹配（$E \leqslant 1 \times 10^{-6}$），Phage-phinder 就可以用 TIGR 的数据库来决定该基因在细菌基因组中的定位，并使用 TIGR 的基因功能分类法（TIGR gene role assignments，Role-ids）[109]帮助识别寄主的蛋白质（表 2）。

表 2　与噬菌体相关的 TIGR Role_ids

基因功能类别	描述
88	细胞外膜，其他
89	胞壁质囊和肽聚糖生物合成
90	表面多糖和脂多糖生物合成与降解
91	表面结构
94	毒素产生与抗性
123	2′-脱氧核糖核酸代谢
129	调控功能，其他
131	DNA 降解
132	DNA 复制、重组和修复
138	蛋白质、肽和糖肽的降解
149	非典型条件的适应
152	原噬菌体功能
154	转座子功能
156	保守的假定蛋白
157	未知功能的普通蛋白
175	普通蛋白
183	限制/修饰系统
185	功能待定
187	致病性
261	调控作用，DNA 相互作用
270	被破坏的可读框架
703	未知特性的酶
704	保守区域

Phage-phinder 沿着基因组扫描，并统计每个窗口中与噬菌体数据库条目匹配的基因数目，为了消除背景干扰，只有找到至少四个匹配窗口才能进一步研究（图 2）。找出该区域最多匹配窗口，并用 TIGR 的 Role-ids 对每个可读框进行检测，直到遇到不感兴趣的 Role-ids（表 2）。图 2 为恶臭假单胞菌 KT2440 基因组检测输出图。

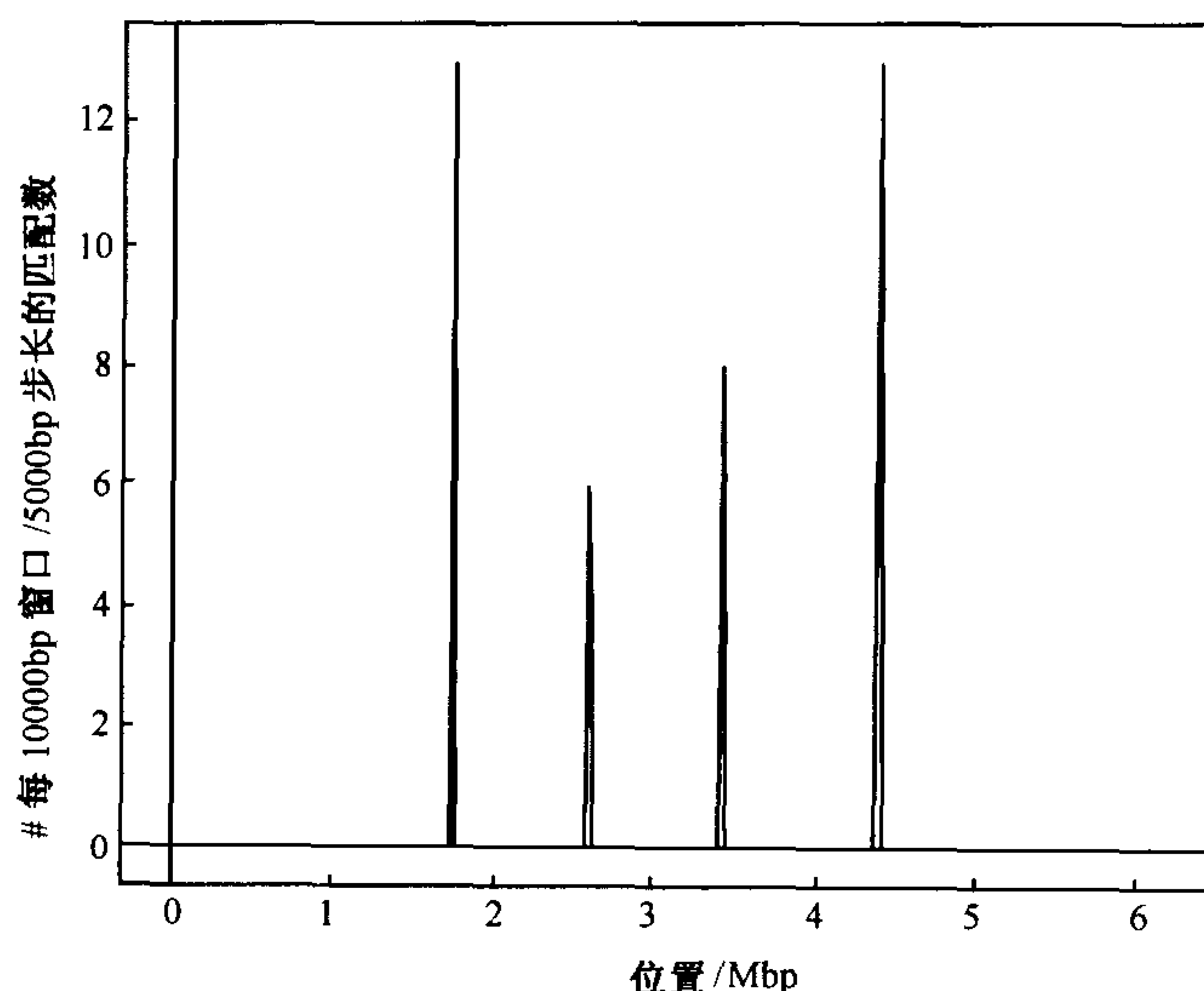


图2　Phage-phinder 原噬菌体定位软件输出图。用 10 000bp 的窗口和 5000bp 的步长分析了恶臭假单胞菌（*Pseudomonas putida*）KT2440 的全基因组[117]。为了消除背景干扰，只有 4 个或更多个窗口标示出来。输出结果为 XGRAPH（General Purpose 2-D Plotter；表 1）的输入文件，该文件先保存为 PostScript 格式而后转换成 tif 文件。

为了弄清楚有多少原噬菌体存在已测序的细菌基因组序列中，以噬菌体数据库为对照，对 CMR 中 90 种细菌菌株中的 89 种进行搜寻，Phage-phinder 程序检测结果表明，所处理的 89 个基因组中，有 141 个碱基数超过 18kbp 假定原噬菌体，总计5 197 329bp，或者说占细菌基因组 DNA 的 1.85%，平均每个基因组含 1.6 种假定原噬菌体。这个数字大于基因库中已完成噬菌体的基因组数目（基因库中有 72 种噬菌体大于 18 kbp，总计 3 129 029bp）（图 3）。当然，它们并不都是功能性噬菌体，但即使有一半是功能性的，仍有约 70 种原噬菌体，占现有噬菌体基因组库的约 80%，这些噬菌体可用于比较和种系发生方面的研究。

目前，还必须根据数据来确定一段序列是否与噬菌体有关，还是只包含了功能性噬菌体的个别基因，当用 Phage-phinder 程序分析另一种细菌基因组时出现了问题，因为 Phage-phinder 发现了不包括其他噬菌体维持基因（phage-keeping gene）的限制修饰基因岛，当 GenBank 中已知序列的噬菌体包含转座元件或 ABC 转运蛋白时，又出现了其他问题，导致许多寄主基因错误地标记在噬菌体区域内。

通过对噬菌体特定蛋白质的准确鉴定，更便于对原噬菌体区域的识别和注释，噬菌体蛋白质家族的极端序列趋异特性可能使 BLAST 搜索无效，用高质量的隐式马可夫模型（HMM）[110,111]能使灵敏度和特异性大大提高。

为了获得建立 HMM 所需多重蛋白序列比对，先需用 BLAST 格式化多重蛋白

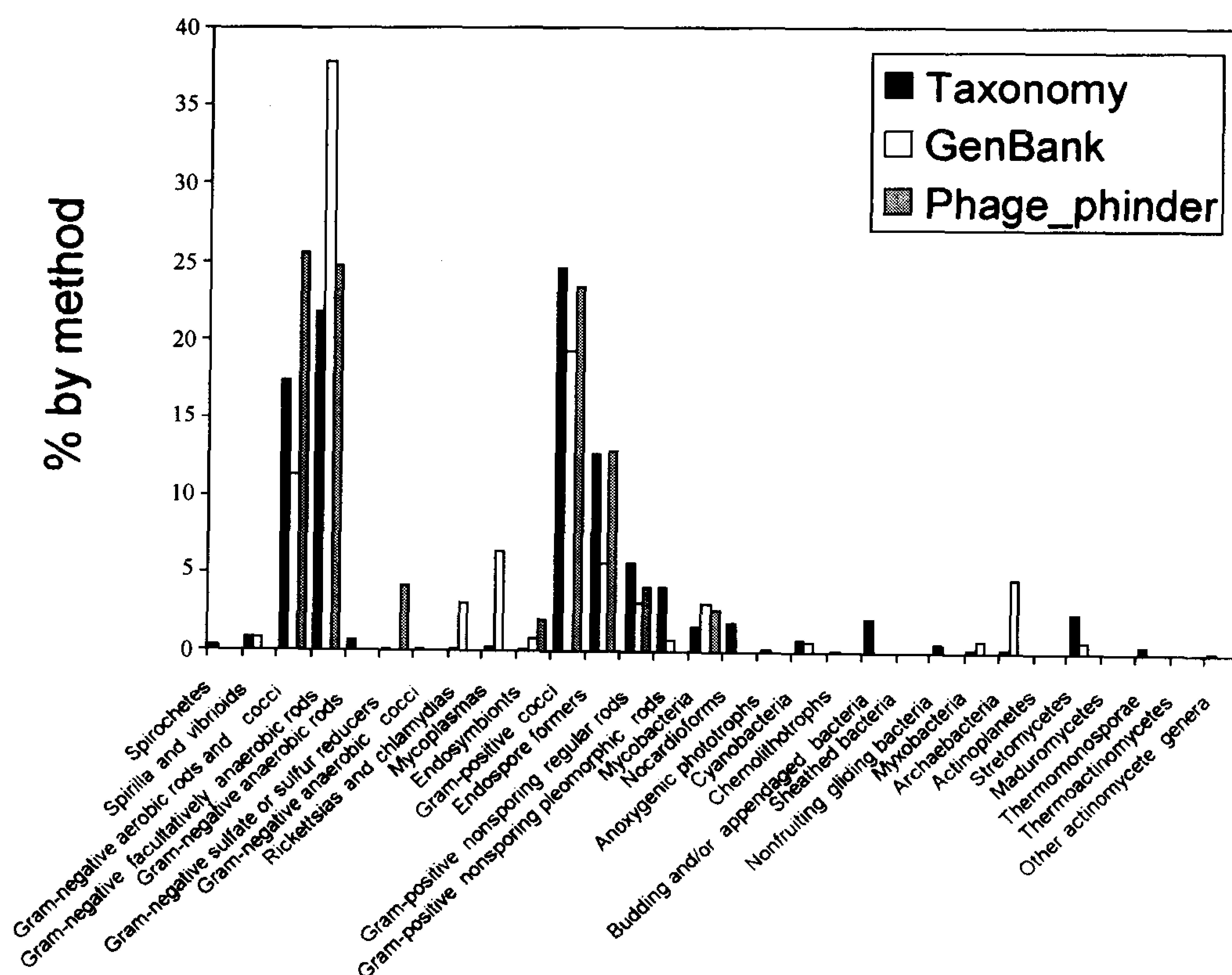


图 3 不同寄主属中噬菌体的数量。细菌分类和分组采用 Ackermann 的方法和伯杰氏细菌鉴定手册[1,123]。噬菌体数量分别用每种方法（Taxonomy、GenBank 和 Phage-phinder）所得噬菌体总数的百分含量表示。Ackermann 的 Taxonomy 数据，只代表 2000 年形态学描述的有尾噬菌体（肌动噬菌体科、管状噬菌体科和足状噬菌体科）数量[1]。GenBank 是 Entrez 基因库中有全序列噬菌体基因组的数目，为 2002 年 9 月 16 日的统计数（表 1）。Phage-phinder 数据来自 89 种全序列细菌基因组和预计大于 18kbp 原噬菌体的分析结果。

FASTA 文件来运行 BLASTP。FASTA 文件含有 Phage-phinder BLASTP 搜索的主题蛋白序列，以及 CMR 中具有推定原噬菌体功能（role-id 152）的所有蛋白。这种“全部对全部（all-vs-all）”的搜索结果由定位分隔形式输出，并输入到产生单链锁簇（single-linkage cluster）的 Perl 语言程序中。

另外创建第二个基于 Perl 的程序，以记数、注解并重新得到含 10 个或更多成员的氨基酸序列簇，包括鉴别出 1793 个氨基酸序列的 82 个簇，其中只有 73 个簇作了进一步分析，因为另外 9 个已经有了相应的 HMM（表 1 和 3）。

表 3 针对噬菌体蛋白质的隐式马可夫模型

PF00959	噬菌体溶解酶
PF01818	噬菌体转录调控蛋白
PF02061	λ 噬菌体 CⅢ
PF02305	衣壳蛋白（F 蛋白）

续表

PF02306	主要的棒状蛋白（G蛋白）
PF02316	Mu DNA结合域
PF02924	λ噬菌体头部修饰蛋白D
PF02925	噬菌体脚手架蛋白D
PF03197	噬菌体FRD2蛋白
PF03245	噬菌体裂解蛋白质
PF03335	噬菌体尾丝重复
PF03354	预测的噬菌体终止酶，大亚基
PF03374	噬菌体反阻抑蛋白
PF03406	噬菌体尾丝重复
PF03420	原头核心蛋白蛋白酶，T4家族
PF03431	RNA复制酶，β链
PF03592	终止酶小亚基
PF03863	噬菌体成熟蛋白
PF03864	噬菌体主要衣壳蛋白E
PF03903	噬菌体T4尾丝
PF03906	噬菌体T7尾丝蛋白
PF04233	噬菌体Mu蛋白F样蛋白
TIGR01446	DnaD和噬菌体相关域
TIGR01537	噬菌体门蛋白，HK97家族
TIGR01538	噬菌体门蛋白，SPP1家族
TIGR01539	噬菌体门蛋白，λ家族
TIGR01541	噬菌体尾部tape measure蛋白，λ家族
TIGR01543	噬菌体原头部蛋白酶，HK97家族
TIGR01547	噬菌体终止酶，大亚基，PBSX家族
TIGR01551	噬菌体主要的衣壳蛋白，P2家族
TIGR01554	噬菌体主要的衣壳蛋白，HK97家族
TIGR01555	噬菌体相关蛋白，HI1409家族
TIGR01558	噬菌体终止酶，小亚基，推定，2P27家族
TIGR01560	特性未知的噬菌体蛋白（可能的DNA包装蛋白）
TIGR01563	噬菌体头-尾接合器，推定
TIGR01592	Holin，SPP1家族
TIGR01593	毒素分泌/噬菌体裂解holin
TIGR01594	噬菌体holin，λ家族
TIGR01598	Holin，噬菌体ΦLC3家族
TIGR01600	噬菌体次尾蛋白L
TIGR01603	噬菌体主尾蛋白，Φ13家族
TIGR01606	Holin，BlyA家族
TIGR01610	噬菌体复制蛋白O，N端区域
TIGR01611	噬菌体主尾管蛋白
TIGR01613	噬菌体/质粒引物酶蛋白，P4家族，C端区域
TIGR01618	噬菌体核酸结合蛋白
TIGR01629	噬菌体/质粒引物酶蛋白，基因Ⅱ/Ⅹ家族
TIGR01630	噬菌体特性未知的蛋白，C端区域
TIGR01633	噬菌体假定的尾部成分，N端区域
TIGR01634	噬菌体尾部蛋白I

续表

TIGR01635	噬菌体病毒粒子形态发生蛋白
TIGR01636	噬菌体转录调控子，RinA 家族
TIGR01637	噬菌体转录调控子，ArpU 家族
TIGR01641	推定的噬菌体头部形态蛋白，SPP1 gp7 家族
TIGR01644	噬菌体基板装配蛋白 V
TIGR01665	噬菌体次结构蛋白，N 端区域
TIGR01669	噬菌体特性未知的蛋白，XkdX 家族
TIGR01671	噬菌体保守假定蛋白 TIGR01671
TIGR01673	噬菌体 holin，LL-H 家族
TIGR01674	噬菌体次尾部蛋白 G
TIGR01712	噬菌体 6-腺苷-甲基转移酶
TIGR01714	噬菌体复制体组织者，假定，N 端区域
TIGR01715	噬菌体尾装配蛋白 T
TIGR01725	噬菌体蛋白，HK97 gp10 家族
TIGR01760	噬菌体 tape measure 蛋白，TP901 家族，核心区域

注：见表 1 查找 Pfam 和 TIGRFAM 网址。

最初，只对那些有尾双链 DNA 噬菌体中，负责包装和头部形成晚期保守基因的基因簇进行了研究[10,13,112]（图 4）。这个保守区域由 5 个或更多编码蛋白的座位组成，并具下列功能：（a）编码 *pac* 或 *cos* 位点的小和大的终止酶亚单位、切割噬菌体基因组串联体，可能涉及噬菌体基因组装配头部[3]；（b）编码门户蛋白形成一个孔或门户以使 DNA 在装配和排出时通过，也可以在噬菌体头部（衣壳）和尾部蛋白间形成连接[113]；（c）编码原头蛋白酶将主衣壳蛋白裂解为成熟形式[112,114]；（d）编码组装噬菌体衣壳的主要衣壳蛋白[112]。

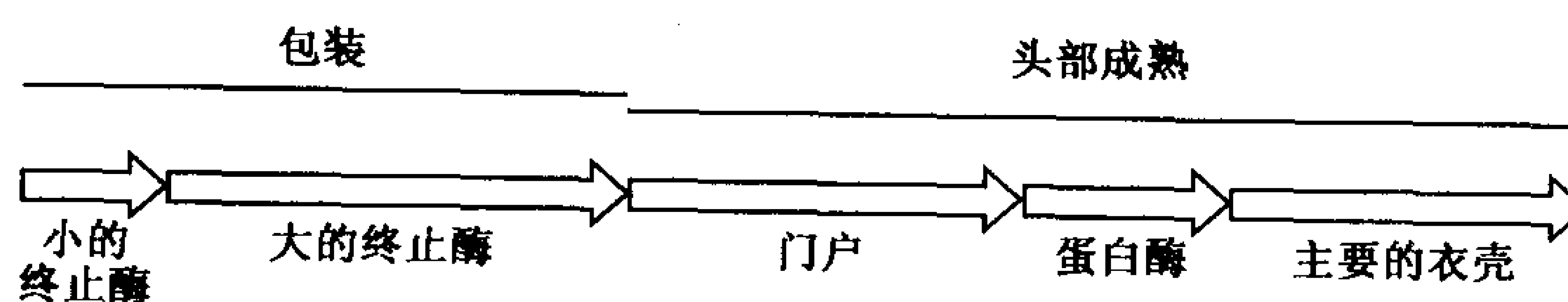


图 4　有尾噬菌体保守装配操纵子和头部成熟操纵子

大部分基因座有多个蛋白序列簇，它们属不同蛋白质家族，执行相似功能，但没有明显的序列相似性。这些簇的已知功能成员可用来进行 PSI-BLAST（Position-Specific Iterated BLAST）搜索（表 1）[115]，以获得更多组的相关蛋白质。每组蛋白质用 CLUSTAL W 排列（表 1）[116]，初步的序列比对经过整修、精选、重排或编辑，以提高序列比对的质量以及在此基础上序列搜索的质量。

这些经过手工加工的序列比对，用 HMMER 软件包进行 HMM 构建和测试（表 1 和表 3）[110]，属于特征明确的家族蛋白容易正确注释，而对较小、变异较大和特征不明确的蛋白质（如 holin、切除酶、尾附属蛋白），事实证明，不用 HMM 要鉴别它们相当难。

Phage-phinder 好不好用呢？起初用 Phage-phinder 测试炭疽芽孢杆菌，采用的窗口

大小为 10 000bp，步长为 5000bp[117]，结果正确识别出 4 个经手工标识的假定原噬菌体（忽略自身标的以避免使结果产生偏差）。在恶臭假单胞菌中发现的 4 个假定原噬菌体区域均被正确标识，然而，在另一细菌基因组中，Phage-phinder 误把大于 18kbp 的一个假定接合转移元件当作原噬菌体区域，但是，它正确识别了所有其他手工标识的原噬菌体区域。

这些假定的原噬菌体区域，必须根据 ICTV（国际病毒分类学委员会）分类系统进行命名和分类（表 1），目前，大多数假定原噬菌体区域是根据 BLAST 搜索结果，用先前已研究清楚的噬菌体来分类并主观命名。现在，由于 Rohwer 和 Edwards[118] 出版了一本突破性专著，噬菌体委员会有了基于基因组的噬菌体系统分类学，他们的方法与存在的分类标准及文献报道一致。

Rohwer 和 Edwards[118] 为了找到能用于建立种系发生树的保守基因或模体（motif），曾搜索 DNA 和蛋白质序列，但没有成功。简言之，确定一个蛋白质组距离（proteomic distance）涉及以下步骤：①采用 CLUSTAL W 为每个 E 值小于 0.1 的 BLASTP 标的（hit）进行序列排比；②采用 PROTDIST 程序（PHYLIP 软件的一部分，Phylogeny Inference Package）确定蛋白质距离，根据蛋白质长度加以修正；③合计所有长度校正后的蛋白质距离，有遗漏蛋白的话，要罚分；④用 PHYLIP 格式产生一个距离矩阵，根据该矩阵可建立种系发生树。Edwards 编的 Perl 程序通过 GNU General Public License 免费提供（表 1）。

这个分析确实支持了许多噬菌体家族现有的分类，但要研究原噬菌体（属于长尾噬菌体科、短尾噬菌体科和肌尾噬菌体科）的特性，却不一定要这么严格。他们提出的命名法由三个基本部分组成：①有几个例子表明，感染不同寄主的噬菌体有相同的名字（如结核分枝杆菌 ΦL5 与酒明串珠菌 ΦL5），所以噬菌体名字由第一个鉴定的寄主属名全称 + 种名全称 + 希腊字母 Φ（表示 phi）+ 噬菌体编号组成。②用 phage 代替 viridae 以表明噬菌体或原噬菌体是用计算机方法进行分类的（如 Siphophage 长尾噬菌体）。③属于特征性亚组噬菌体或原噬菌体的命名，用 like 一词附在该亚组研究最清楚的亚群噬菌体名上（如 Φ29-like）。

结束语

鉴定和注释原噬菌体区域的软件需要改进，因为已经有大量未被使用的公共信息可以获得，这对噬菌体的比较遗传学、群体生物学及进化生物学的发展具有重要作用。需要建立更多的 HMM 模型，尤其是对那些未被充分代表的噬菌体基因和那些小的、高度趋异和在配对排列搜索时容易被遗漏的基因。对原噬菌体区域更好地鉴定及注释，将会大大有效地提高对细菌基因组的注释，因为多达 7% 细菌基因组可能源于噬菌体[119]。

来自肠细菌的噬菌体基因组序列，以及来自奶工业中制造酸奶和干酪的细菌噬菌体基因组序列，在公共数据库中有充分的代表。还有更多细菌噬菌体基因组需要测序，首先要从那些已经具有丰富生物学信息的噬菌体开始。令人烦恼的是，帮助弄清了转化原理的噬菌体 T1 还没有完全测序，而且也未登录在基因库中。还有很多特性已弄清的噬菌体还没有完全测序和登录在基因库中（例如，许多化脓链球菌噬菌体 A24、T12 和

C1)。正如我们所担忧的，一些命了名的噬菌体，还不在基因库中或序列不完整或没有独立条目，如噬菌体 T2（感染大肠杆菌菌株 B)、CTFΦ（编码霍乱毒素的丝状霍乱弧菌噬菌体)[55]、SopEΦ（P2 -样噬菌体，感染伤寒沙门氏菌、携带有通过 III 型分泌系统分泌的毒性因子 SopE 的[65]，炭疽芽孢杆菌的 γ 噬菌体[120]和鼠伤寒沙门氏菌的 *Gifsy*噬菌体[62,121,122]。具有重要历史和医药意义但还没有完全测序和登录的噬菌体太多了，以致无法在这章中完全列出，这迫切要求噬菌体研究者共同对它们进行测序，并将其完整基因组登录，就像 20 世纪 40 年代那样协同努力。

所有这些噬菌体和原噬菌体的基因组信息，需要登录在由因特网可以公开得到的数据库中，尽管基因库是一个很大的资源，但并不是所有的完整噬菌体序列都被列入噬菌体的 Entrez 基因组里（表 1），即使列入基因库中，也有寄主不清楚、注释前后不一致或者没有注释。如果来自细菌基因的原噬菌体，包含在一个专门详细组织的资源里，那么，噬菌体之间的比较将会更加方便。

由于大多数测序的细菌基因组都是动物或植物的病原体，全球有大约 10^{30}噬菌体的生物多样性是至关重要的[89]，必须对从未能培养的环境样品（海水、河水、受污染场所或其他小生境）中得到的噬菌体进行测序，在这样的环境中可能存在很多的生物多样性，这将有助于理解存在自然界中的噬菌体种类以及它们所携带的基因。众所周知，噬菌体能携带和转移细菌毒性和适应性的基因[7,51]，是否有一些有助于生物整理的基因呢？噬菌体已帮助我们深入了解了分子机制，谁又知道什么生物秘密隐藏于未知微生物的噬菌体基因组中呢？

致谢

感谢 Karen E. Nelson 邀请我来撰写本书的一章，感谢 Martin Wu 和 Brian Haas 帮助整理 BLASTP btab 数据，感谢 Dan Haft 建立 HMM、Scott Durkin 关于噬菌体注释的讨论、Jonathan Eisen 帮助阐述蛋白质距离测定、Robert（Bob）Koenig 和 Aymeric de Vazeille 帮助把德文翻译成英文、Bebbie Rhodes 出色的图书资料支持、Jacques Ravel、Garry Myers 和 Eddy Arnold-Berkovitz 帮助转换绘图文件，还要感谢 Tim Read 对我在基因组研究所从事噬菌体生物信息学的兴趣和努力工作所给予不断而诚恳的鼓励和支持。

（刘作易，刘明秋　译）

参考文献

1. Ackermann HW. Frequency of morphological phage descriptions in the year 2000. Brief review. Arch Virol 2001; 146:843–857.
2. Maniloff J, Ackermann HW. Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. Arch Virol 1998; 143:2051–2063.
3. Ackermann HW. Tailed bacteriophages: the order *Caudovirales*. Adv Virus Res 1999; 51: 135–201.
4. Campbell A. Phage evolution and speciation. In: Calendar R (ed). The Bacteriophages. New York: Plenum, 1988, pp. 1–14.

5. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA 1999; 96:2192–2197.
6. Hendrix RW. Evolution: the long evolutionary reach of viruses. Curr Biol 1999; 9:R914–R917.
7. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. The origins and ongoing evolution of viruses. Trends Microbiol 2000; 8:504–508.
8. Hendrix RW. Bacteriophages: evolution of the majority. Theor Popul Biol 2002; 61:471–480.
9. Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M-C. Analysis of six prophages in *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage populations. Nucleic Acids Res 2001; 29:644–651.
10. Desiere F, Lucchini S, Brüssow H. Comparative sequence analysis of the DNA packaging, head, and tail morphogenesis modules in the temperate cos-site *Streptococcus thermophilus* bacteriophage Sfi21. Virology 1999; 260:244–253.
11. Lucchini S, Desiere F, Brüssow H. Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. J Virol 1999; 73:8647–8656.
12. Desiere F, Mahanivong C, Hillier AJ, Chandry PS, Davidson BE, Brüssow H. Comparative genomics of lactococcal phages: insight from the complete genome sequence of *Lactococcus lactis* phage BK5-T. Virology 2001; 283:240–252.
13. Desiere F, McShan WM, van Sinderen D, Ferretti JJ, Brüssow H. Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic streptococci: evolutionary implications for prophage–host interactions. Virology 2001; 288:325–341.
14. Brüssow H, Desiere F. Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. Mol Microbiol 2001; 39:213–222.
15. Meijer WJ, Horcajadas JA, Salas M. ϕ29 family of phages. Microbiol Mol Biol Rev 2001; 65: 261–287.
16. Weisberg RA, Gottesmann ME, Hendrix RW, Little JW. Family values in the age of genomics: comparative analyses of temperate bacteriophage HK022. Annu Rev Genet 1999; 33:565–602.
17. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 2000; 299:27–51.
18. Brüssow H. Phages of dairy bacteria. Annu Rev Microbiol 2001; 55:283–303.
19. Ackermann HW, Martin M, Vieu JF, Nicolle P. Felix d'Hérelle: His life and work and the foundation of a bacteriophage reference center. ASM News 1982; 48:346–348.
20. Duckworth DH. History and basic properties of bacterial viruses. In: Goyal SM, Gerba CP, Bitton G (eds). Phage Ecology. New York: Wiley, 1987, pp. 1–43.
21. d'Hérelle F. La coccobacille des sauterelles. Ann Inst Pasteur Paris 1914; 28:280–328.
22. d'Hérelle F. Sur un microbe invisible antagoniste des bacilles dysentériques. CR Acad Sci Paris 1917; 165:373–375.
23. Twort FW. An investigation on the nature of ultra-microscopic viruses. Lancet 1915; 2:1241–1243.
24. Ruska H. Die Sichtbarmachung der Bakteriophagen Lyse im Übermikroskop. Naturwiss 1940; 28:45–46.
25. Ruska H. Über ein neues bei der Bakteriophagen Lyse auftretendes Formelement. Naturwiss 1941; 29:367–368.
26. Pfankuch E, Kausche GA. Isolierung und übermikroskopische Abbildung eines Bakteriophages. Naturwiss 1940; 28:46.
27. Luria SE, Anderson TF. The identification and characterization of bacteriophages with the electron microscope. Proc Natl Acad Sci USA 1942; 28:127–130.
28. Anderson TF. Electron microscopy of phages. In: Cairns J, Stent GS, Watson JD (eds). Phage and the Origins of Molecular Biology. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1992, pp. 63–78.

29. Luria SE, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 1943; 28:491–511.
30. Hershey AD, Chase M. Independent functions of viral protein and nucleic acids in growth of bacteriophage. J Gen Physiol 1952; 36:39–56.
31. Zinder ND, Lederberg J. Genetic exchange in *Salmonella*. J Bacteriol 1952; 64:679–699.
32. Benzer S. On the topography of genetic fine structure. Proc Natl Acad Sci USA 1961; 47: 403–415.
33. Crick FHC, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. Nature (London) 1961; 192:1227–1232.
34. Terzaghi E, Okada Y, Streisinger G, Emrich J, Inouye M, Tsugita A. Change of a sequence of amino acids in phage T4 lysozyme by acridine-induced mutations. Proc Natl Acad Sci USA 1966; 56:500–507.
35. Linn S, Arber W. Host specificity of DNA produced by *Escherichia coli*, X. In vitro restriction of phage fd replicative form. Proc Natl Acad Sci USA 1968; 59:1300–1306.
36. Weiss B, Richardson CC. Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from *Escherichia coli* infected with T4 bacteriophage. Proc Natl Acad Sci USA 1967; 57:1021–1028.
37. Weiss B, Jacquemin-Sablon A, Live TR, Fareed GC, Richardson CC. Enzymatic breakage and joining of deoxyribonucleic acid. VI. Further purification and properties of polynucleotide ligase from *Escherichia coli* infected with bacteriophage T4. J Biol Chem 1968; 243:4543–4555.
38. Silber R, Malathi VG, Hurwitz J. Purification and properties of bacteriophage T4-induced RNA ligase. Proc Natl Acad Sci USA 1972; 69:3009–3013.
39. De Waard A, Paul AV, Lehman IR. The structural gene for deoxyribonucleic acid polymerase in bacteriophages T4 and T5. Proc Natl Acad Sci USA 1965; 54:1241–1248.
40. Panet A, van de Sande JH, Loewen PC, et al. Physical characterization and simultaneous purification of bacteriophage T4 induced polynucleotide kinase, polynucleotide ligase, and deoxyribonucleic acid polymerase. Biochemistry 1973; 12:5045–5050.
41. Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science 1985; 228:1315–1317.
42. Benson N, Sugiono P, Bass S, Mendelman LV, Youderian P. General selection for specific DNA-binding activities. Genetics 1986;114:1–14.
43. MacWilliams MP, Celander DW, Gardner JF. Direct genetic selection for a specific RNA-protein interaction. Nucleic Acids Res 1993; 21:5754–5760.
44. Salser W, Fry K, Brunk C, Poon R. Nucleotide sequencing of DNA: preliminary characterization of the products of specific cleavages at guanine, cytosine, or adenine residues (bacteriophage M13–ribosubstitution–DNA polymerase I–electrophoresis–two-dimensional fingerprinting). Proc Natl Acad Sci USA 1972; 69:238–242.
45. Schreier PH, Cortese R. A fast and simple method for sequencing DNA cloned in the single-stranded bacteriophage M13. J Mol Biol 1979; 129:169–172.
46. Kunkel TA. Rapid and efficient site-specific mutagenesis without phenotypic selection. Proc Natl Acad Sci USA 1985; 82:488–492.
47. Kunkel TA, Roberts JD, Zakour RA. Rapid and efficient site-specific mutagenesis without phenotypic selection. Methods Enzymol 1987; 154:367–382.
48. Chauthaiwale VM, Therwath A, Deshpande VV. Bacteriophage lambda as a cloning vector. Microbiol Rev 1992; 56:577–591.
49. Lwoff A, Horne R, Tournier P. A system of viruses. Cold Spring Harbor Symp Quant Biol 1962; 27:51–55.
50. Frobisher M, Brown J. Transmissible toxicogenicity of streptococci. Bull. Johns Hopkins Hosp 1927; 41:167–173.

51. Wagner PL, Waldor MK. Bacteriophage control of bacterial virulence. Infect Immun 2002; 70: 3985–3993.
52. Fujii N, Oguma K, Yokosawa N, Kimura K, Tsuzuki K. Characterization of bacteriophage nucleic acids obtained from *Clostridium botulinum* types C and D. Appl Environ Microbiol 1988; 54:69–73.
53. Holmes RK, Barksdale L. Genetic analysis of tox+ and tox– bacteriophages of *Corynebacterium diphtheriae*. J Virol 1969; 3:586–598.
54. Uchida T, Gill DM, Pappenheimer AM. Mutation in the structural gene for diphtheria toxin carried by temperate phage β. Nat New Biol 1971; 233:8–11.
55. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 1996; 272:1910–1914.
56. Huang A, de Grandis S, Friesen J, et al. Cloning and expression of the genes specifying Shiga-like toxin production in *Escherichia coli* H19. J Bacteriol 1986; 166:375–379.
57. Recktenwald J, Schmidt H. The nucleotide sequence of Shiga toxin (Stx) 2e-encoding phage ϕP27 is not related to other Stx phage genomes, but the modular genetic structure is conserved. Infect Immun 2002; 70:1896–1908.
58. Betley MJ, Mekalanos JJ. Staphylococcal enterotoxin A is encoded by phage. Science 1985; 229:185–187.
59. Bensing BA, Rubens CE, Sullam PM. Genetic loci of *Streptococcus mitis* that mediate binding to human platelets. Infect Immun 2001; 69:1373–1380.
60. Wright A. Mechanism of conversion of the *Salmonella* O antigen by bacteriophage ε^{34}. J Bacteriol 1971; 105:927–936.
61. Guan S, Bastin DA, Verma NK. Functional analysis of the O antigen glucosylation gene cluster of *Shigella flexneri* bacteriophage SfX. Microbiology 1999; 145(Pt 5):1263–1273.
62. Figueroa-Bossi N, Bossi L. Inducible prophages contribute to *Salmonella* virulence in mice. Mol Microbiol 1999; 33:167–176.
63. Hynes WL, Ferretti JJ. Sequence analysis and expression in *Escherichia coli* of the hyaluronidase gene of *Streptococcus pyogenes* bacteriophage H4489A. Infect Immun 1989; 57:533–539.
64. Sako T, Sawaki S, Sakurai T, Ito S, Yoshizawa Y, Kondo I. Cloning and expression of the staphylokinase gene of *Staphylococcus aureus* in *Escherichia coli*. Mol Gen Genet 1983; 190: 271–277.
65. Mirold S, Rabsch W, Rohde M, et al. Isolation of a temperate bacteriophage encoding the type Ⅲ effector protein SopE from an epidemic *Salmonella typhimurium* strain. Proc Natl Acad Sci USA 1999; 96:9845–9850.
66. Fischetti VA. Phage antibacterials make a comeback. Nat Biotechnol 2001; 19:734–5.
67. Nelson D, Loomis L, Fischetti VA. Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. Proc Natl Acad Sci USA 2001; 98:4107–4112.
68. Loeffler JM, Nelson D, Fischetti VA. Rapid killing of *Streptococcus pneumoniae* with a bacteriophage cell wall hydrolase. Science 2001; 294:2170–2172.
69. Stone R. Bacteriophage therapy. Stalin's forgotten cure. Science 2002; 298:728–731.
70. Summers WC. Bacteriophage therapy. Annu Rev Microbiol 2001; 55:437–451.
71. Biswas B, Adhya S, Washart P, et al. Bacteriophage therapy rescues mice bacteremic from a clinical isolate of vancomycin-resistant *Enterococcus faecium*. Infect Immun 2002; 70:204–210.
72. Chernomordik AB. Bacteriophages and their therapeutic-prophylactic use. Med Sestra 1989; 48:44–47.
73. Schuch R, Nelson D, Fischetti VA. A bacteriolytic agent that detects and kills *Bacillus anthracis*. Nature 2002; 418:884–889.

74. Le Y, Gagneten S, Tombaccini D, Bethke B, Sauer B. Nuclear targeting determinants of the phage P1 cre DNA recombinase. Nucleic Acids Res 1999; 27:4703–4709.
75. Groth AC, Olivares EC, Thyagarajan B, Calos MP. A phage integrase directs efficient site-specific integration in human cells. Proc Natl Acad Sci USA 2000; 97:5995–6000.
76. Schagen FH, Rademaker HJ, Cramer SJ, et al. Towards integrating vectors for gene therapy: expression of functional bacteriophage MuA and MuB proteins in mammalian cells. Nucleic Acids Res 2000; 28:E104.
77. Olivares EC, Hollis RP, Calos MP. Phage R4 integrase mediates site-specific integration in human cells. Gene 2001; 278:167–176.
78. Stoll SM, Ginsburg DS, Calos MP. Phage TP901-1 site-specific integrase functions in human cells. J Bacteriol 2002; 184:3657–3663.
79. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
80. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 1977; 265:687–695.
81. Sanger F, Coulson AR, Friedmann T, et al. The nucleotide sequence of bacteriophage phiX174. J Mol Biol 1978; 125:225–246.
82. Daniels DL, Sanger F, Coulson AR. Features of bacteriophage lambda: analysis of the complete nucleotide sequence. Cold Spring Harb Symp Quant Biol 1983; 47(Pt 2):1009–1024.
83. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. J Mol Biol 1982; 162:729–773.
84. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 1997; 390:249–256.
85. Glaser P, Frangeul L, Buchrieser C, et al. Comparative genomics of *Listeria* species. Science 2001; 294:849–852.
86. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 2001; 409:529–533.
87. Campbell AM. Cryptic prophages. In: Neidhardt FC, Curtiss R Ⅲ, Ingraham JL, et al. (eds). *Escherichia coli* and *Salmonella* Cellular and Molecular Biology. Washington, DC: American Society for Microbiology Press, 1996, pp. 2041–2046.
88. Read TD, Brunham RC, Shen C, et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 2000; 28:1397–1406.
89. Brüssow H, Hendrix RW. Phage genomics: small is beautiful. Cell 2002; 108:13–16.
90. Ferretti JJ, McShan WM, Ajdic D, et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 2001; 98:4658–4663.
91. Pečenková T, Benes V, Pačes J, Vlček C, Pačes V. Bacteriophage B103: complete DNA sequence of its genome and relationship to other *Bacillus* phages. Gene 1997; 199:157–163.
92. Osborn MA, Böltner D. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid 2002; 48:202–212.
93. Nakayama K, Takashima K, Ishihara H, et al. The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage. Mol Microbiol 2000; 38:213–231.
94. Uratani Y, Hoshino T. Pyocin R1 inhibits active transport in *Pseudomonas aeruginosa* and depolarizes membrane potential. J Bacteriol 1984; 157:632–636.
95. Daw MA, Falkiner FR. Bacteriocins: nature, function and structure. Micron 1996; 27:467–479.
96. Rampling A, Whitby JL. Preparation of phage-free pyocin extracts for use in the typing of *Pseudomonas aeruginosa*. J Med Microbiol 1972; 5:305–312.
97. Jones LF, Zakanycz JP, Thomas ET, Farmer JJ 3rd. Pyocin typing of *Pseudomonas aeruginosa*: a simplified method. Appl Microbiol 1974; 27:400–406.

98. Kurtz S, Schleiermacher C. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 1999; 15:426–427.
99. Zhao S, Williams KP. Integrative genetic element that reverses the usual target gene orientation. J Bacteriol 2002; 184:859–860.
100. Morgan GJ, Hatfull GF, Casjens S, Hendrix RW. Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. J Mol Biol 2002; 317:337–359.
101. Masignani V, Giuliani MM, Tettelin H, Comanducci M, Rappuoli R, Scarlato V. Mu-like prophage in serogroup B *Neisseria meningitidis* coding for surface-exposed antigens. Infect Immun 2001; 69:2580–2588.
102. Zimmer M, Scherer S, Loessner MJ. Genomic analysis of *Clostridium perfringens* bacteriophage ϕ3626, which integrates into *guaA* and possibly affects sporulation. J Bacteriol 2002; 184:4359–4368.
103. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res 2002; 30:866–875.
104. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997; 25:955–964.
105. Shimizu-Kadota M, Kiwaki M, Sawaki S, Shirasawa Y, Shibahara-Sone H, Sako T. Insertion of bacteriophage phiFSW into the chromosome of *Lactobacillus casei* strain Shirota (S-1): characterization of the attachment sites and the integrase gene. Gene 2000; 249:127–134.
106. Nelson KE, Weinel C, Paulsen IT, et al. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ Microbiol 2002; 4:799–808.
107. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The Comprehensive Microbial Resource. Nucleic Acids Res 2001; 29:123–125.
108. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol 1990; 215:403–410.
109. Riley M. Functions of the gene products of *Escherichia coli*. Microbiol Rev 1993; 57:862–952.
110. Eddy SR. Profile hidden Markov models. Bioinformatics 1998; 14:755–63.
111. Wang IN, Smith DL, Young R. Holins: the protein clocks of bacteriophage infections. Annu Rev Microbiol 2000; 54:799–825.
112. Duda RL, Martincic K, Hendrix RW. Genetic basis of bacteriophage HK97 prohead assembly. J Mol Biol 1995; 247:636–647.
113. Casjens S, Hendrix R. Control mechanisms in dsDNA bacteriophage assembly. In: Calendar R (ed). The Bacteriophages. New York: Plenum Press, 1988, pp. 15–91.
114. Black LW, Showe MK, Steven AC. Morphogenesis of the T4 head. In: Karam J (ed). Molecular Biology of Bacteriophage T4. Washington, DC: American Society for Microbiology Press, 1994, pp. 518–558.
115. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
116. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673–4680.
117. Takami H, Nakasone K, Takaki Y, et al. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 2000; 28:4317–4331.
118. Rohwer F, Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 2002; 184:4529–4535.
119. Simpson AJ, Reinach FC, Arruda P, et al. The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature 2000; 406:151–157.

120. Brown ER, Cherry W. Specific identification of *Bacillus anthracis* by means of a variant bacteriophage. J Infect Dis 1955; 96:34–39.
121. Figueroa-Bossi N, Coissac E, Netter P, Bossi L. Unsuspected prophage-like elements in *Salmonella typhimurium*. Mol Microbiol 1997; 25:161–173.
122. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. Mol Microbiol 2001; 39: 260–271.
123. Holt JG (ed). Bergey's Manual of Systematic Bacteriology. Baltimore, MD: Williams & Wilkins, 1984.

第三部分：核心功能

6

微生物代谢比较

Karen E. Nelson

引言

1995年，流感嗜血菌（*Haemophilus influenzae*）[1]基因组测序完成并公诸于世，从此，开启了微生物基因组研究的新领域。迄今为止，已有100多种微生物基因组完成测序并公布序列，估计全世界还有300种以上微生物测序正在进行（www.tigr.org）。在微生物基因组发展的早期阶段，优先被测序的显然是具有医学重要价值的物种，如流感嗜血菌[1]和生殖道支原体（*Mycoplasma genitalium*）[2]。毫无疑问，基因组测序有助于对主要病原菌基础生物学的认识，以进一步遴选更多抗微生物的靶标[3]。

基因组测序重点随后逐渐转向其他领域，如农业、环境、进化和生物技术[4]。基因组测序除了解释病原菌可能的致病机制[5~7]，还能进一步认识微生物物种之间的水平基因转移[8]，近缘种之间的基因组重排[7,9~11]，以及一些微生物种群在环境中的潜在利用价值[12]。微生物全基因组测序为了解未知微生物个体、群体以至群落（community）的基本生化信息打通了一条新路。

本章将着力介绍基因组测序如何增进对微生物基础生物学、生物化学和生理学的了解，以及如何将微生物基因组内和基因组之间进行比较，以获取新测序物种未知的生物学信息。文中列举的例子是已发表的文献和正在进行的基因组研究资料。

挖掘微生物基因组序列中的信息

鉴于其他章节已阐述了构建基因组文库和测定微生物全基因组序列的方法学，本章不再赘叙有关技术细节。一旦基因组组装完成，所有物理缺口和测序缺口都被封闭，就必须用生物信息学分析技术解释该物种的生物学问题。生物信息学分析可以识别所有的可读框（ORF），以及其他特征，如基因组中的tRNA、rRNA、重复序列等，这些分析还可能延伸到鉴定基因间区域、碱基偏爱性、复制起点、潜在基因水平转移区域、插入序列元件和质粒等。

有效的基因自动识别可以用隐藏式马可夫模型（hidden Markov model, HMM）[13~15]完成，也可以用内插式马可夫模型，如基因定位（Gene Locator）和内插式马可夫模型（Interpolated Markov Modeler）（Glimmer[14]）这样的软件来完成，并将计算机程序自动化与人工管理相结合来确定基因生物意义的类别和功能。BLAST（Basic Local Alignment Search Tool）[16,17]或FASTA可在序列数据库中搜寻和比较蛋白质同源族，包括HMM、Pfam和COG（Clusters of Orthologous Group，直系同源群簇），帮助

进行功能预测。基因组注解的更多细节，详见第 3 章。

仔细分析基因组序列和对文献的详尽研究，可以重新勾画出某物种的生理学轮廓，并预测未知的生化代谢途径。在多数情况下，首轮注解和人工管理尚不能解决代谢途径的所有步骤，但是，通过设制 HMM 模型，或者与其他物种相似代谢途径的序列做比较和搜寻未知物种的同源序列，往往可以填补代谢途径中的空白。需要强调的是，即使一个基因组注解完后，还会有相当多的 ORF 被作为保守假定蛋白（conserved hypothetical protein）（与其他物种有同源性），或假定蛋白（hypothetical protein）（为该物种特有）。由于目前分析方法和对微生物多样性认识上的局限，无疑尚需对在物种中具有生物学作用又不能确定在细胞中功能的 ORF 作进一步研究。

尽管有可能从各种途径获得许多生化途径的代谢图谱，但没什么软件直接从基因组数据库中构建微生物生理文档。可以查看代谢图的数据库有：KEGG（Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg/kegg2.html），WIT（What Is There, http://wit.mcs.anl.gov/WIT2/），以及 EcoCyc（http://www.ecocyc.org/）。

KEGG 数据库以基因组信息为基础，主要为物种的利用提供生物信息资源。截至 2003 年，KEGG 数据库中有 132 个物种的信息，代表 481 325 个基因和 5415 种化学反应。

WIT 数据库力图以各个物种的基因序列、生化及表型特征为基础重建代谢模型。目前已有 25 个处于不同阶段的在建项目，对已经完成生化重建的物种，其表示方式为显示基因组中已鉴定出的代谢途径的一个清单和存列 ORF 及其对应功能的一个表。

EcoCyc 数据库基本上以大肠杆菌基因组序列为基础，致力于建立大肠杆菌的生化途径，它已经成为大肠杆菌及其有共同生化途径相似物种的电子文献来源。在 EcoCyc 使用的途径/基因组浏览器（Pathway/Genome Navigator）中，在用户界面上用户可以直观地看到大肠杆菌基因和相当数量的生化反应途径。还提供大肠杆菌的许多其他信息，如已知的所有信号转导途径、操纵子、启动子、转录因子及转录因子结合位点。EcoCyc 是以电子形式描述单个物种基因网络关系包含信息最完整的一个数据库。

EMP 数据库（Enzymes and Metabolic Pathways, http://www.empproject.com/）是整合了酶学和代谢诸多方面生化数据的综合资源，它包括约 15 000 条支持数据库中任何描述有实验数据的发表参考文献。它包含的信息有：细胞培养条件、酶及其反应、酶动力学、热力学以及物理化学，还有 3000 多个代谢图谱，用于代谢网络和代谢设计的数学模拟。

此外，明尼苏达大学生物催化与生物降解数据库（the University of Minnesota Biocatalysis/Biodegradation Database, UM-BBD; http://umbbd.ahc.umn.edu/）主要是一个关于微生物对异生物物质（xenobiotic compound）和化学合成物质（chemical compound）的生物催化和降解途径的电子资源。因此，它提供的信息基本上都与生物技术领域中的重要反应有关，包括生化反应和代谢途径中的起始和中间化合物、能转化化合物的微生物种类、有关酶和基因。

至今，还没有哪个工具能在基因组数据基础之上单独自动或可靠地重建细胞的全部生理过程，主要原因是目前所有代谢分析工具都是依赖基因名称而不是实际序列，这

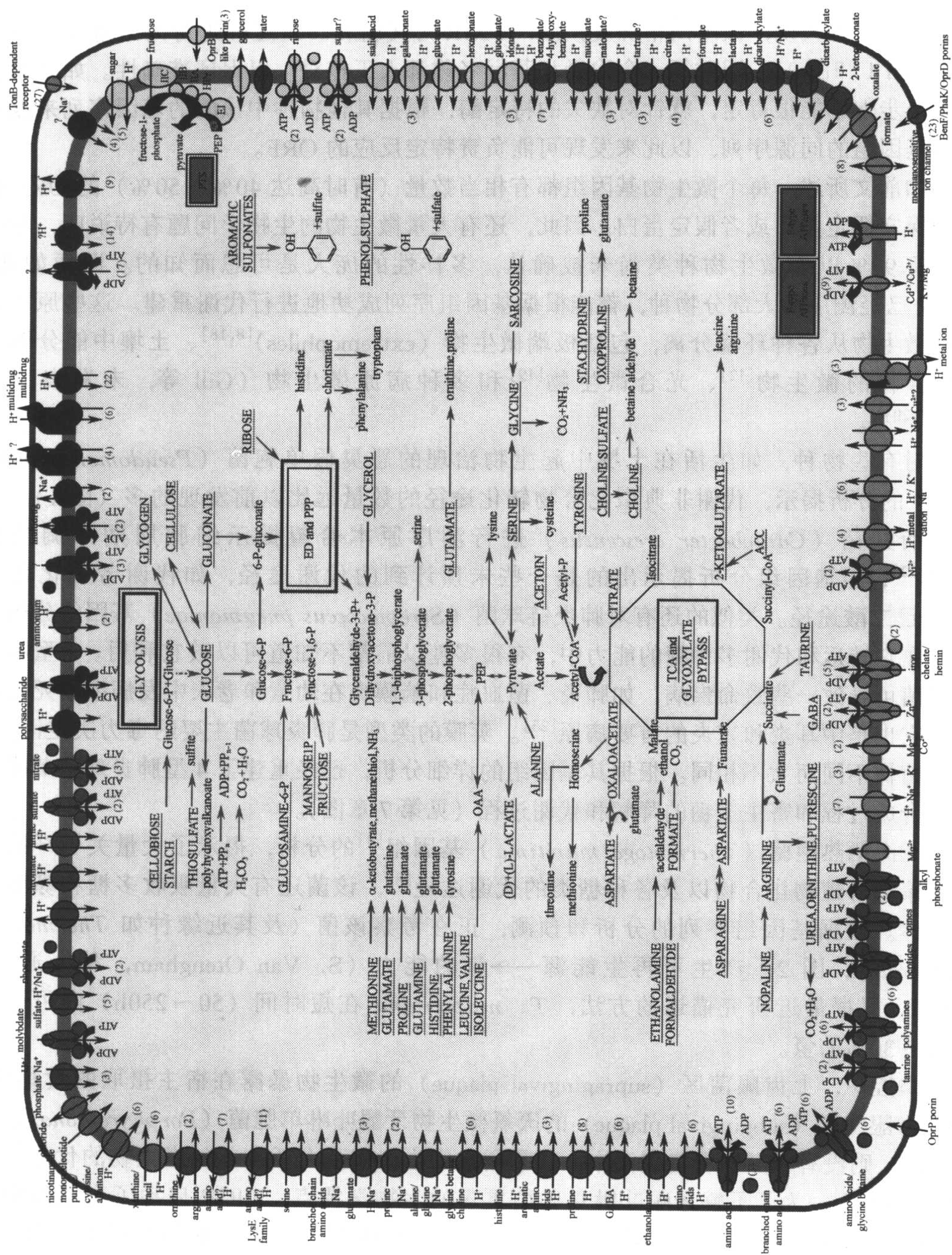


图1　根据恶臭假单胞菌（*Pseudomonas putida*）基因组信息重建的代谢途径。复制经《环境微生物学》同意。Nelson KE，Weinel C，Paulsen IT. et al. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environmental Microbiology 2003，5（7）：630.（另见文前彩色插图 6-1）

样，可传递的错误（transitive error）（见第 3 章）就被整合到最终的代谢途径预测中。为了减轻这种失误，做出高精确度预测，TIGR 对代谢和转运过程进行人工重建（见第 7 章，图 1 和图 2），即对每一个途径和基因名称都人工核对，以保证准确性。如果途径中某个步骤不能很肯定，就针对缺失的特定酶，根据其他物种中该酶的基因序列来搜寻待测基因组的同源序列，以此来发现可能负责特定反应的 ORF。

如前文所述，每个微生物基因组都有相当数量（有时高达 40%～50%）的 ORF 被定为保守假定蛋白或者假定蛋白。因此，还有大量微生物的生物学问题有待说明，考虑到还有 99% 以上微生物种类尚未被确认，多样性的庞大是可想而知的。即便如此，TIGR 已经测序的大部分物种，都能根据基因组序列成功地进行代谢重建。这些原核和真核微生物从各种环境分离，包括极端微生物（extremophiles）[8,18]、土壤中能分解芳香化合物的微生物[12]、光合微生物[19]和多种病原微生物（Gill 等，未发表的手稿[5,7,20]）。

对有些物种，如生活在土壤中起生物治理的恶臭假单胞菌（*Pseudomonas putida*）[12]的分析揭示，代谢非典型化合物转化途径的数量远比以前发现的多（图 1）。对新月柄杆菌（*Caulobacter crescentus*）进行测序原本希望揭示细胞周期的调控机制[21,22]，但基因组分析揭示出的是一些未预计到的代谢途径，如代谢芳香化合物 β-酮-己二酸途径。类似的还有对肺炎链球菌（*Streptococcus pneumoniae*）基因组分析发现，在它转运和代谢多种糖的能力中，有很多糖以前都不知道可以被它利用。4 型肺炎链球菌可导致一些致命疾病，如肺炎、菌血症和脑膜炎在幼童和老人中发病率和死亡率高，它也是中耳炎和窦炎的首要病原[23]。荚膜的类型是肺炎球菌主要的毒力决定因子，并因菌株不同而大不相同。根据其基因组的详细分析，已经重建了 4 型肺炎球菌荚膜的生物合成过程和寄主多糖的降解和代谢过程（见第 7 章图）。

对海栖热袍菌（*Thermotoga maritima*）基因组[8]的分析，揭示了大量关于纤维素和木聚糖等植物化合物以及各种糖类的代谢途径[8]，该菌还有大量吸收多糖和寡肽的转运体。根据基因组序列的分析和预测，正在考察该菌（及其近缘种如 *Thermotoga neopolitana*，图 2）产生可再生能源——氢的能力（S. Van Otengham，私人通讯，2003）。根据最近研究描述的方法，*T. neapolitana* 在短时间（50～250h）内可产生 20%～30% 的氢。

口腔中，上齿龈菌区（supragingival plaque）的微生物暴露在宿主摄取的食物中，而下齿龈菌区（subgingival plaque）的厌氧微生物牙龈卟啉单胞菌（*Porphyromonas gingivalis*）则没有暴露在食物残渣中，却暴露在寄主组织蛋白质和其他微生物的代谢终产物环境中[24]。为了了解牙龈卟啉单胞菌更多的生物学特点，TIGR 启动了全基因组计划，并于 2002 年完成[25]。尽管已经知道牙龈卟啉单胞菌对葡萄糖的利用度很低[26]，但测序发现它其实并不含有糖酵解途径中所有酶的 ORF。而且，还发现一些可能的葡萄糖/半乳糖转运体和葡萄糖激酶的 ORF。当找到了磷酸戊糖途径中的 4 个假定 ORF 时，推测该菌可能通过该途径在厌氧环境下产生代谢前体。天冬氨酸、天冬酰胺和谷氨酰胺都可被牙龈卟啉单胞菌利用[24,26]，而从基因组推断出的途径暗示，还有其他氨基酸可以被利用。总共鉴定出 44 个肽酶和一些降解复杂氨基糖的酶。

基因组分析的结果表明，牙龈卟啉单胞菌的主要发酵终产物，包括丙酸、丁酸、异

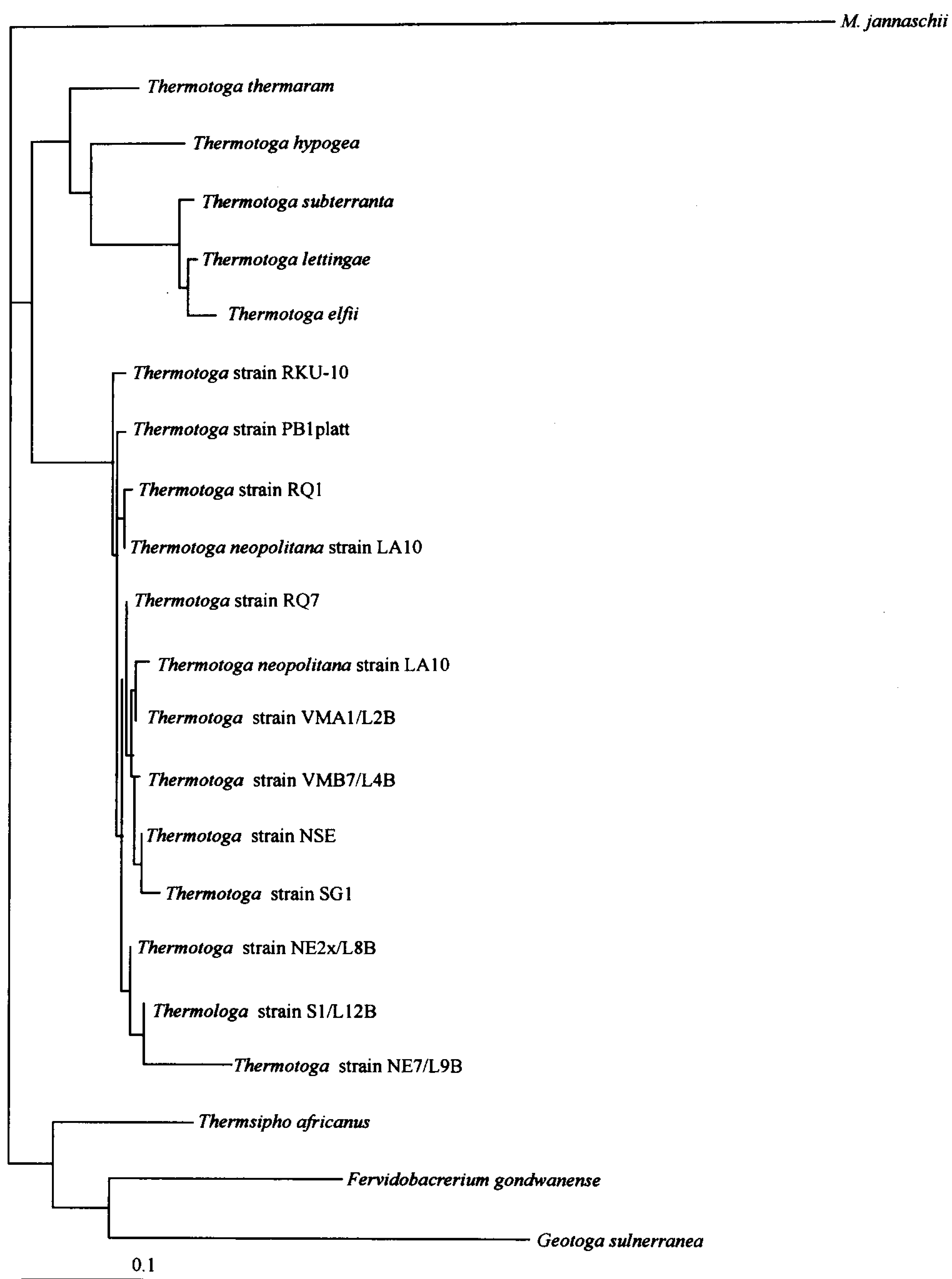


图 2 栖热袍菌目（*Thermotogales*）的系统分类

丁酸、异戊酸、乙酸、乙醇和丁醇，其中一部分对寄主组织有毒，可能导致细胞死亡、免疫细胞活性和细胞素网络受损[27,28]。因此，这些终产物可能是损伤人体寄主的毒力因子。

真核生物举例：恶性疟原虫

作为全基因组分析的另一个例子，对恶性疟原虫（*Plasmodium falciparum*）生化途径的重建[29]（图 3）能更深入地理解其生物学，也能尝试性地确定新药靶标。从基因组序列看，所有糖酵解途径中的必需酶以及磷酸戊糖途径所有酶的候选基因（除一个基因尚未确定外）都可以确定。磷酸烯醇式丙酮酸羧化酶的存在，表明恶性疟原虫可能通过这个酶从细胞质中的磷酸烯醇式丙酮酸和重碳酸盐反应来补充草酰乙酸，以弥补 TCA 循环对中间产物草酰乙酸的消耗。

生物化学、遗传学和化学治疗数据表明，疟疾及其他顶复门寄生虫能够通过莽草酸途径从 4-磷酸赤藓糖和磷酸烯醇式丙酮酸合成分支酸（chorismate）[30,31]。除分支酸合成酶外，还不能从基因组序列中肯定莽草酸途径中其他酶的基因。

最后，疟原虫利用寄主细胞质中的血红蛋白作为食物来源，水解珠蛋白产生血红素，后者以疟色素（hemazoin）的形式被解毒。还不明确开始合成时是利用转运的宿主酶，还是利用寄生虫自身的酶，从基因组序列分析看，代谢途径中除了尿卟啉原-III 合成酶外，其他每一个酶的直系同源物都能找到。

微生物培养研究中的基因组学

从基因组信息可以预测出微生物的底物利用模式，这对重建微生物代谢途径有重要价值。例如，耐辐射异常球菌（*Deinococcus radiodurans*）的全基因组序列[32]，可以帮助确定在营养受限制的辐射环境中保持其生长的关键营养要素[33]。为了发展一种基本合成培养基，可试图以不同含量组合的碳源、氨基酸、盐和维生素的液体和固体培养基，这种方法能确保基本生长所必需的营养成分以及浓度。耐辐射异常球菌的生长依赖于可代谢的碳源、外源氨基酸和维生素，而满足其生长特别需要富硫氨基酸和烟酸。由于在多种氨基酸复合物中都能生长，而没有哪种氨基酸复合物是生长特别必需的。影响生长好坏的主要因素是培养基中的总氨基酸浓度，而不是氨基酸复合物的组成。可以预计，该技术将会应用到大量正在测序而培养仍然困难或不能培养的微生物，来确定其培养条件，例如产乙烯脱卤拟球菌（*Dehalococcoides ethenogenes*），目前还需要微生物提取物来支持其生长，又如 *Epulopiscium* 菌目前还不能在实验室培养。

Epu 是锯尾鲷某些种的肠内共生体。这些异养生物是已知的最大细菌之一，长度能达到 600μm[34]。尽管人们对这个物种已渐有所知，但它们仍不能在实验室中培养。种系发生学分析表明它们是属于低 G+C 含量的革兰氏阳性菌[34]。

该细菌的一个特异之处是它能够在体内产生多胞胎（multiple offspring），这种子细胞产生方式可能代表了细胞繁殖的一种新形态进化的下个步骤[35]。尽管仍不能在实验室中培养，Epu 的基因组序列还是可以提供一些关于它在体内产生多胞胎的重要信息，以及原核细胞向真核细胞过渡的过程中发生过的一些早期变化。此外通过对 Epu 基因组的功能注释，有可能对其进行代谢重建并成功培养。

出于对这个物种的极大兴趣，基因组研究所（TIGR）最近启动了 Epu 的基因组测

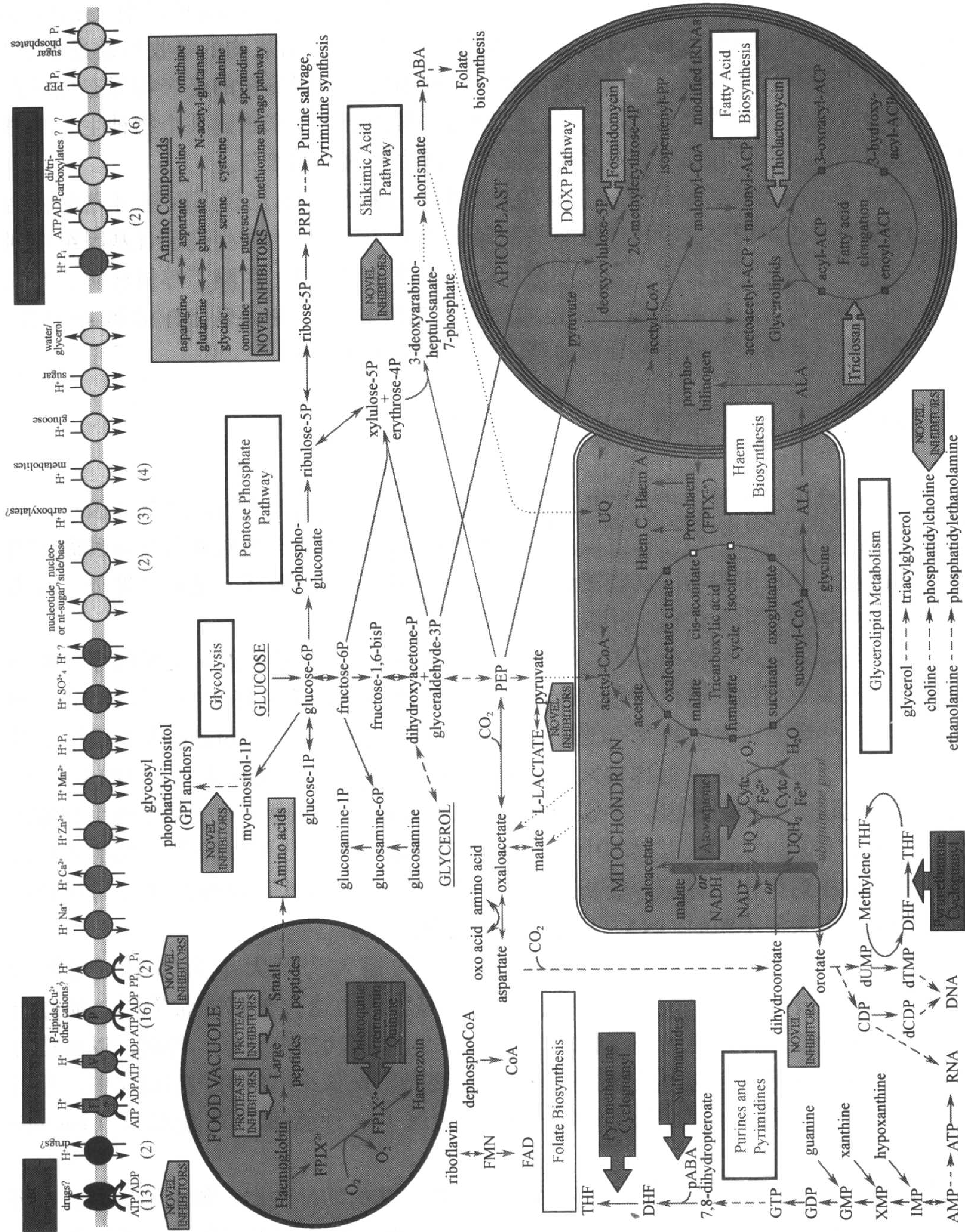


图 3 依照恶性疟原虫基因组而重建的代谢途径图。经《自然》杂志允许而重印。Gardner M J, Hall N, Fung E, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002; 419 (6906): 498-511.（另见文前彩色插图 6-3）

序计划，其目的在于解析这个物种中存在的主要代谢途径并由此鉴定出一批候选底物，使人们可以在实验室中培养这种不同寻常的物种。测序所需的部分 DNA 由以噬菌体

φ29DNA 多聚酶为工具，从细胞或噬菌斑中直接进行的 DNA 多引物滚环扩增（multiply primed rolling circle amplification of DNA）所提供[36,37]。Epu 对底物的吸收可通过在 C^{14} 标记的底物中培养样品来检测。底物从培养基中的消失以及待测菌的终产物形态也可由气相色谱—质谱分析来确定。可以想像，这种通过基因组测序来重建某个物种生长培养基的方法同样也可用于其他为数众多的未能培养微生物的成功培养。

除了对 Epu 的底物利用分析，与那些产内生孢子的细菌的比较基因组研究也将会开展，以鉴定这些物种中与子代产生有关的保守基因。这些产生内生孢子且基因组序列可用的细菌包括枯草芽孢杆菌、炭疽芽孢杆菌、耐盐芽孢杆菌、丙酮丁醇梭杆菌、产气荚膜梭菌。艰难梭菌、肉毒梭菌、破伤风梭菌、热纤梭菌、梭状芽孢杆菌属 sp BC1、枯草芽孢杆菌以及嗜热脂肪芽孢杆菌的基因组测序也即将完成。

用比较代谢预测法理解生态

已获得的基因组数据，代表了几百种几乎都是能纯培养的微生物，这些基因组信息所反映的是有限的物种多样性和生理多样性。实际上，天然生态系统通常由各种微生物群落之间的相互作用而驱动，这些微生物承担了生物地理化学循环、元素再循环、污染物的降解、生物量转化等过程，并维系着人类与动物的健康。

现在，可以通过对环境中 DNA 基因组（metagenomic，泛基因组）文库测序，来破译多样性生态系中物种的遗传信息。这种方法的关键在于对某种环境样本构建 BAC 和 fosmid 文库，然后对文库中的大片段基因组 DNA 进行正确分析，DeLong 及其同事成功地将该技术用于海洋浮游生物样本，该研究组创造性地对样本文库大片段 DNA 测序，以确定未经纯培养物种的基因。在他们系列研究报告的第一篇中，对一种浮游生物海洋古生菌构建了 fosmid 文库，并对其中一个 40kb 基因组片段进行了分析[38]。这类大片段携带的核糖体序列可作为标记，分析这些有种属特征的片段就可以得出相关的物种信息，如果只分析来源不明的 DNA，则很难得到结果。

DeLong 及其同事随后构建了 BAC 文库，平均插入片段大小为 80kb，最大为 155kb[39]，研究表明，BAC 文库对于提供未经培养物种的遗传信息非常有用，通过分析这些大插入片段，可以提供基因组成、基因组织和基因功能等信息。同时，以这些片段为线索可确定在种系上与未经培养物种相近的物种。在该小组的另一篇论文中，通过分析以前未培养过的一种物种的 BAC 文库，发现了一种新细菌视紫红质（bacteriorhodopsin）[40]。

这种技术已经用到分析其他环境中的大量未能培养的物种样本，如土壤、人的口腔和胃肠道、马尾藻海（Sargasso Sea），以便了解这些微生物的更多信息和这些天然种群的基因组潜能。一般情况下，分析这些物种遗传能力的基本技术是：过滤和回收样本中所有的原核细胞，构建小的、中度和大的文库，获得达到预先确定覆盖率要求的足够序列。关键是成功组装所有序列数据，以便产生那些含操纵子或种系发生标记的连续 DNA 片段。最理想的是完全组装出这些环境样本中以前未能培养微生物的全基因组，这还要取决于组装软件能否胜任这一挑战。

以人为例，人依赖于寄居在人体内部和表面各种小生境中数十亿微生物的活动，尽

管其中有很多微生物与寄主之间处于动态和互惠关系，但也存在一些能引起慢性感染乃至致死疾病的机会病原物。而研究微生物-寄主相互关系的困难在于，绝大多数微生物难以在实验室培养，其鉴定和描述要依赖于分子手段。

显然，数量庞大的未定性微生物对人体的健康和疾病的影响是巨大的，而这种影响还有待阐明。许多慢性病，如克罗恩氏病（Crohn's disease，节段性回肠炎）和川崎病（Kawasaki's disease），有感染性疾病的特点，却没有微生物病原存在的有力证据。对这些微生物种群的鉴定和描述，无疑将能建立它们与感染和慢性病之间相联结的关系，并阐明它们在免疫系统发育过程中的作用，以及在整体上对人类进化的影响[41]。通过全面研究人体微生物及其基因组，来阐明这些微生物在人体环境中的作用还有漫长的路要走。

口腔微生物对维持人体口腔健康和引起牙周病至关重要。全面考察健康个体和患病个体体内的可培养和不可培养微生物及其基因组多样性有助于了解如下方面：第一，可以帮助鉴定口腔内生活的所有细菌和古生菌种类；第二，可以鉴定种群中编码毒力因子或涉及口腔生物膜形成、菌落定殖（colonization）和对杀菌药物产生抗性的基因。大量新基因的发现终将有助于开发控制口腔和全身中由多种微生物种群引起的感染性疾病。总之，在口腔生物膜中由多种复杂微生物组成的小生境，对防止外来细菌种类扩增和保持寄主口腔健康起着重要作用。

采用 16S rDNA 的分类和鉴定方法发现了不同环境样本中大量细菌和古生菌，然而，这种方法只能评价微生物物种多样性和存在的种群，却不能深入了解种群内的遗传多样性。在这些环境中，有哪些基因是以前从全基因组测序或在公共数据库的其他序列中从未出现过的呢？同样，16S rDNA 分类法也不能确定个体在全新环境中的生理过程，然而，通过 16S rDNA 序列比较可能得到物种之间的某些亲缘关系，但是，已经越来越多地认识到从 16S rDNA 序列分析鉴定出的近缘物种在基因组组成上可能存在极大差异[42]。由于这些微生物不能培养，而对从环境样本中直接获得的这些大片段 DNA 序列进行测序和分析就更有意义了。深入理解这些生态系统中的遗传信息以及它们之间的相互作用，将有助于了解生态群落是如何影响系统运作的整个过程。如果得到环境样本的全部遗传信息，就可以通过表达分析手段监测群落是如何应对不同环境压力而做出适应性变化的。

可以预见，通过详尽的生物信息分析，可以用类似注解单个微生物基因组序列的方法处理整个群落，同样，可以构建所有预测 ORF 及其假定注解，以及生物学功能目录的数据库，通过研究所有功能，就可以确定在指定环境中的所有生化途径。像 Beja 及其同事通过对一个 fosmid 克隆的研究确定了一种视紫红质[40]那样，如果对复杂环境样本的物种测序，将可获得大量生物化学和物理化学数据。可以想像，假如对土壤泛基因组进行分析，可发现许多代谢环式化合物/芳香化合物和新抗生素，以及次级代谢产物的合成途径。如果分析人胃肠道泛基因组，便可发现大量关于胃肠道寄居微生物代谢能力、发酵能力和代谢终产物的信息。这些发现又反过来能更深入地阐明人胃肠道功能、人体健康以及这些微生物的致病能力。将健康个体的泛基因组与患病个体（即胃肠道功能失常性疾病）的泛基因组进行比较，则可能发现致病微生物物种和致病因子。

从基因组序列分析微生物代谢的前景

物种的生理全过程除了以基因组序列为基础应用生物信息学来阐明外，还可通过基因组功能研究技术，如微阵列和蛋白质组技术等方法获得其细节。DNA 微阵列可以确定不同生长条件下基因的表达情况，而比较基因组杂交（comparative genome hybridization, CGH）则可用来研究近缘物种间基因组的多样性。

目前，可以获得微生物在不同生长条件下基因表达调控的数据，例如，已经研究了多杀巴斯德菌（*Pasterurella multocida*）在营养缺陷条件下的基因表达[43]，用微阵列技术比较了在丰富培养基和基本培养基中的基因表达，669 个基因的表达模式可检测出其变化，大部分变化的基因在丰富培养基中能高水平表达。在基本培养基中表达的上调基因涉及氨基酸生物合成和转运系统、外膜蛋白和热休克蛋白。微阵列技术还用于研究脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）血清型 B 与人上皮细胞相互作用时的基因表达[44]，研究发现，寄主和细菌之间的相互作用可诱导 347 个基因的表达，上调基因包括负责铁、氯化物、氨基酸和硫酸盐转运的蛋白质、许多毒力因子和含硫氨基酸的所有代谢途径。

微阵列还成功地用于检测混合微生物种群环境样本中的基因表达。Dennis 及其同事构建了一个微阵列芯片，并在2,4-D(2,4-二氯苯氧基乙酸）降解细菌富营养罗斯通氏菌（*Ralstonia eutropha*）的纯培养物和混合培养物中加入诱导剂2,4-D，然后将培养物的总 RNA 用微阵列分析[45]。在总细胞浓度为 10^8 个细胞/ml 的混合培养物中，即使 *Ralstonia eutropa* 细胞浓度低于 10^5 个细胞/ml，仍可检测到 5 个2,4-D 代谢基因中的 2 个基因表达量发生了变化，并成功地证明了用微阵列技术能研究复杂生态系统中的基因表达。

Taroncher-Oldenburg 及其同事[46]开发了一种微阵列方法，可对氮循环中的基因表达进行定量检测，对 Choptank 河/切萨皮克海湾区（Chesapeake Bay area）两处的泥沙样品进行检测呈现出不同的杂交结果，这表明，在上述环境梯度（environmental gradient）里，脱氮菌群的组成有非常大的差异，再次证明基因芯片技术应用环境微生物学中的优势。

同样，比较基因组杂交还成功地用于确定近缘种之间的代谢能力差异。对海栖热袍菌（*T. maritima*）全基因组的分析显示，该菌能与同一环境中的其他微生物发生广泛的基因转移，随后用兼并引物 PCR（degenerate PCR primer）扩增和消减杂交（subtractive hybridization）等方法，也证实了基因组多样性和属内基因交换的广泛存在[47,48]。

Nesbo 及其同事[47]用栖热袍菌属的 16 个菌株和栖热袍菌目的其他相近菌株进行了研究，他们在全基因组序列分析的基础上，对众多预测为“类古生菌”基因中 2 个基因的分布进行了深入研究。这两个基因编码谷氨酸合成酶大亚基和肌醇-1-磷酸合成酶，它们的分布模式表明来源于栖热袍菌目进化树中的多个古生菌枝系，而与其他细菌种类无关。

在后续研究中，Nesbo 及其同事[48]用抑制消减杂交技术（suppressive subtractive hybridization）确定，在不同栖热袍菌属菌株中存在的那些与已测序栖热袍菌属菌株基因组没有同源基因。重点研究了 *Thermotoga* sp RQ2 已测序海栖热袍菌 MSB8 的 16S

rRNA 基因只有 0.3% 差异，基因有大量差异。

在 TIGR，我们对由德国雷根斯堡大学（University of Regensburg）Karl Stetter 和 Robert Huber 两博士赠送、自世界各地分离的栖热袍菌属菌株用比较基因组杂交方法加以研究，以期更深入了解关于该属内普遍存在基因转移现象的详情。初步结果显示，海栖热袍菌 MSB8 与 *Thermotoga* sp RQ2 的基因组有高度保守性，其 rRNA 小亚基序列 99.7% 相似。MSB8 基因组与 RQ2 至少有 7%（129 个 ORF）没有同源序列，在 129 个 ORF 中，有 45 个假定蛋白，13 个保守假定蛋白，23 个（占 18%）转运蛋白。只有 18 个是单一 ORF，其余为 2～38kb 的 DNA 片段，其中大多都编码 RQ2 中没有的糖代谢途径。RQ2 菌株缺乏塔格糖、果胶、核糖和甘油的代谢途径。比较基因组杂交研究确定了分离自不同地点的各个菌株中主要的操纵子差异（即操纵子的存在与否），这些差异与分离地点相关，这或许反映出在不同地点的不同底物可利用性存在差异。

抑制消减杂交技术和微阵列技术可以联用，以确定不同生态系统中的微生物种群差别，例如比较健康人和患者的消化道菌群。White 及其同事已经成功地运用抑制消减杂交技术确定了不同饲料喂养牛的瘤胃内微生物种群的差别（B. White，个人通讯，2003 年 8 月）。这也反映出，瘤胃内微生物种群的差异与瘤胃所摄入食物的差异相适应。

单个微生物和整个环境的代谢建模

Schilling 及其同事[49]在利用基因组和生物化学数据预测各种微生物代谢网络方面成绩斐然。他们认为，在缺乏动力学常数的情况下对细胞行为的预测能力是有限的，同时也认为，即使在缺乏动力学信息的情况下，仍有可能建立完整细胞过程，如代谢理论模型，并有可能在稳态假设下研究可能的代谢流分布。稳态分析对代谢网络有一些假定限制条件，基于物料平衡假定条件的代谢网络稳态分析，亦称为代谢流平衡分析（flux balance analysis，FBA）。在系列研究的一篇论文中，他们利用现有文献，注解后的基因组序列数据和菌种特异性信息，构建出代表大肠杆菌 MG1655 全细胞范围内代谢能力的计算机模型。然后，再用 FBA 对这些限制条件下的代谢能力进行评估，从而对该条件下细胞的生长情况作定性预测，这些预测最终会对分析混合生态系中微生物的生物化学反应起极为重要的作用。

展望

在微生物个体水平，基因组中接近 40% 的序列为假定蛋白或保守假定蛋白，目前已获得的大量数据要求开发高通量方法来更有效地分析这些数据，包括高通量蛋白质组学、基因表达和蛋白质-蛋白质相互作用研究。预期这些方法能进一步延伸应用到微生物群落的分析中，而要研究微生物群落之间复杂的相互关系，还需开发新技术手段，以评价已获得培养和未能培养（估计占微生物种类中 99% 以上）物种之间的联系。前文提到，微阵列正迅速成为实验室的一项用来研究不同条件下的基因表达，以及与参照基因组比较，不同菌株或物种间基因存在或缺失的标准技术。微阵列的应用扩展到微生物研究的许多方面，包括微生物生理学、发病机制、流行病学、生态学、种系发生和代谢

途径工程。

最近的研究表明，通过弄清微生物的营养条件并模拟天然环境中各营养成分的浓度，一些以前不能培养的物种可以成功培养。Zengler 及其同事报道了一种高通量培养方法，用这种方法可以在低营养流条件下对包裹在凝胶微滴中的微囊化细胞进行微培养，通过流式细胞仪检测出那些含有微菌落的凝胶微滴[50]，这种技术还可以成功地用于多种环境。尽管从未能培养的物种中获得遗传材料已经取得一些进展，但如果要进一步认识这些物种，最终仍然需要对它们进行培养。以前不能培养的物种在培养技术上取得进展，就意味着新（基因组）数据将不断产生，而且需要把这些（基因组）数据转化为具微生物生化和生理意义的信息。很显然，这些生化和生理学的预测需要得到大量的证实，而像 Biolog Plates^{TM}（Haywood, CA）这样的现有技术也需要进一步改进，才能用来测试从全基因组分析中得出的新底物。

尽管基因组学在微生物学研究领域取得了巨大进展，但仍然没有一种原核生物的全部基因功能得到完整阐释，这个事实与本章列出的其他挑战将激励人们继续应用基因组学方法去探索微生物代谢之谜。

（欧阳立明，刘　超　译）

参考文献

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
2. Fraser CM, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995; 270: 397–403.
3. Hoffman SL, Rogers WO, Carucci DJ, Venter JC. From genomics to vaccines: malaria as a model system. Nat Med 1998; 4:1351–1353.
4. Nelson KE, Paulsen IT, Heidelberg JF, Fraser CM. Status of genome projects for nonpathogenic bacteria and archaea. Nat Biotechnol 2000; 18:1049–1054.
5. Tettelin H, Saunders NJ, Heidelberg S, et al: Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
6. Tettelin H, Masignani V, Cieslewicz MJ, et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci USA 2002; 99:12,391–12,396.
7. Tettelin H, Nelson, KE, Pausen IT, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 2001; 293:498–506.
8. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 1999; 399:323–329.
9. Read TD, et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 2000; 28:1397–1406.
10. Read TD, Myers GS, Brunham RC, et al. Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. Nucleic Acids Res 2003; 31:2134–2147.
11. Paulsen IT, Seshadri R, Nelson KE, et al. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci USA 2002; 99: 13,148–13,153.

12. Nelson KE, Weinel C, Paulsen T, et al. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ Microbiol 2002; 4:799–808; erratum in Environ Microbiol 2003; 5:630.
13. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with Glimmer. Nucleic Acids Res 1999; 27:4636–4641.
14. Eddy SR. Noncoding RNA genes. Curr Opin Genet Dev 1999; 9:695–699.
15. Henderson J, Salzberg S, Fasman KH. Finding genes in DNA with a hidden Markov model. J Comput Biol 1997; 4:127–141.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403–410.
17. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
18. Klenk HP, Clayton RA, Tomb JF, et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 1997; 390:364–370.
19. Eisen JA, Nelson KE, Paulsen IT, et al. The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci USA 2002; 99: 9509–9514.
20. Tomb JF, White O, Kerlavage AR, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997; 388:539–547.
21. Nierman WC, Feldblyum N, Laub MT, et al. Complete genome sequence of *Caulobacter crescentus*. Proc Natl Acad Sci USA 2001; 98:4136–4141.
22. Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L. Global analysis of the genetic network controlling a bacterial cell cycle. Science 2000; 290:2144–2148.
23. Paton JC, Andrew PW, Boulnois GJ, Mitchell TJ. Molecular analysis of the pathogenicity of *Streptococcus pneumoniae*: the role of pneumococcal proteins. Annu Rev Microbiol 1993; 47: 89–115.
24. Shah HN, Williams RAD. Utilization of glucose and amino acids by *Bacteroides intermedius* and *Bacteroides gingivalis*. Curr Microbiol 1987; 15:241–246.
25. Nelson KE, Fleischmann RD, DeBoy RT, et al. The complete genome sequence of the oral pathogenic bacterium *Porphyromonas gingivalis* strain W83. J. Bacteriol 2003; 185:5591–5601.
26. Takahashi N, Sato T, Yamada T. Metabolic pathways for cytotoxic end product formation from glutamate- and aspartate-containing peptides by *Porphyromonas gingivalis*. J Bacteriol 2000; 182:4704–4710.
27. Niederman R, Brunkhorst B, Smith S, Weinreb RN, Ryder MI. Ammonia as a potential mediator of adult human periodontal infection: inhibition of neutrophil function. Arch Oral Biol 1990; 35(Suppl):205S–209S.
28. Niederman R, Zhang J, Kashket S. Short-chain carboxylic-acid-stimulated, PMN-mediated gingival inflammation. Crit Rev Oral Biol Med 1997; 8:269–290.
29. Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002; 419:498–511.
30. Roberts CW, Roberts I, Lyons RE, et al. The shikimate pathway and its branches in apicomplexan parasites. J Infect Dis 2002; 185(Suppl 1):S25–S36.
31. Roberts F, Roberts CW, Johnson JJ, et al. Evidence for the shikimate pathway in apicomplexan parasites. Nature 1998; 393:801–805.
32. White O, Eisen JA, Heidelberg JF, et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 1999; 286:1571–1577.
33. Venkateswaran A, McFarlan SC, Ghosal D, et al. Physiologic determinants of radiation resistance in *Deinococcus radiodurans*. Appl Environ Microbiol 2000; 66:2620–2626.
34. Angert ER, Clements KD, Pace NR. The largest bacterium. Nature 1993; 362:239–241.

35. Angert ER, Losick RM. Propagation by sporulation in the guinea pig symbiont *Metabacterium polyspora*. Proc Natl Acad Sci USA 1998; 95:10,218–10,223.
36. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res 2001; 11:1095–1099.
37. Dean FB, Hosono S, Fang L, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA 2002; 99:5261–5266.
38. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bacteriol 1996; 178:591–599.
39. Beja O, Suzuki MT, Koonin EV, et al. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ Microbiol 2000; 2:516–529.
40. Beja O, Aranind L, Koenin EV, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 2000; 289:1902–1906.
41. Relman DA, Falkow S. The meaning and impact of the human genome sequence for microbiology. Trends Microbiol 2001; 9:206–208.
42. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 2001; 409:529–533.
43. Paustian ML, May BJ, Kapur V. Transcriptional response of *Pasteurella multocida* to nutrient limitation. J Bacteriol 2002; 184:3734–3739.
44. Grifantini R, Bartolini E, Muzzi A, et al. Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. Nat Biotechnol 2002; 20:914–921.
45. Dennis P, Edwards EA, Liss SN, Fulthorpe R. Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl Environ Microbiol 2003; 69:769–778.
46. Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB. Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl Environ Microbiol 2003; 69:1159–1171.
47. Nesbo CL, L'Haridon S, Stetter KO, Doolittle WF. Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol Biol Evol 2001; 18:362–375.
48. Nesbo CL, Nelson KE, Doolittle WF. Suppressive subtractive hybridization detects extensive genomic diversity in *Thermotoga maritima*. J Bacteriol 2002; 184:4475–4488.
49. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palson BO, et al. Genome-scale metabolic model of *Helicobacter pylori* 26695. J Bacteriol 2002; 184:4582–4593.
50. Zengler K, Toledo G, Rappe M, et al. Cultivating the uncultured. Proc Natl Acad Sci USA 2002; 99:15,681–15,686.

7

膜转运蛋白的基因组学分析

Ian T. Paulsen, Katherine H. Kang, Mark E. Hance, and Qinghu Ren

引言

细胞膜是细胞的通透性屏障，膜转运蛋白对底物的跨膜运输至关重要。转运蛋白具备的功能有：有机养分的吸收、有毒化合物的排出、离子动态平衡的保持、环境的应激性、能量的产生以及其他重要的细胞功能。据预测，在细菌基因组中有3%～15%的可读框（ORF）编码膜转运蛋白[1]，这进一步突出了转运蛋白在细胞生活方式中的重要性。

膜转运蛋白通过几种不同的能量来源耦合机制介导可溶物质的转运（图1[2]）。初级主动运输转运蛋白利用化学能和光能，通过ATP这种最常用的能源来驱动运输，次级主动运输转运蛋白利用质子、钠及其盐离子或浓度梯度等形式的化学渗透能来驱动转运，膜通道蛋白允许小分子物质或离子自由扩散进入细胞。磷酸转移酶系统在转运的同时将其糖类底物磷酸化，在磷酸化过程中，磷酸烯醇式丙酮酸既是能量供体，也是磷酸供体。

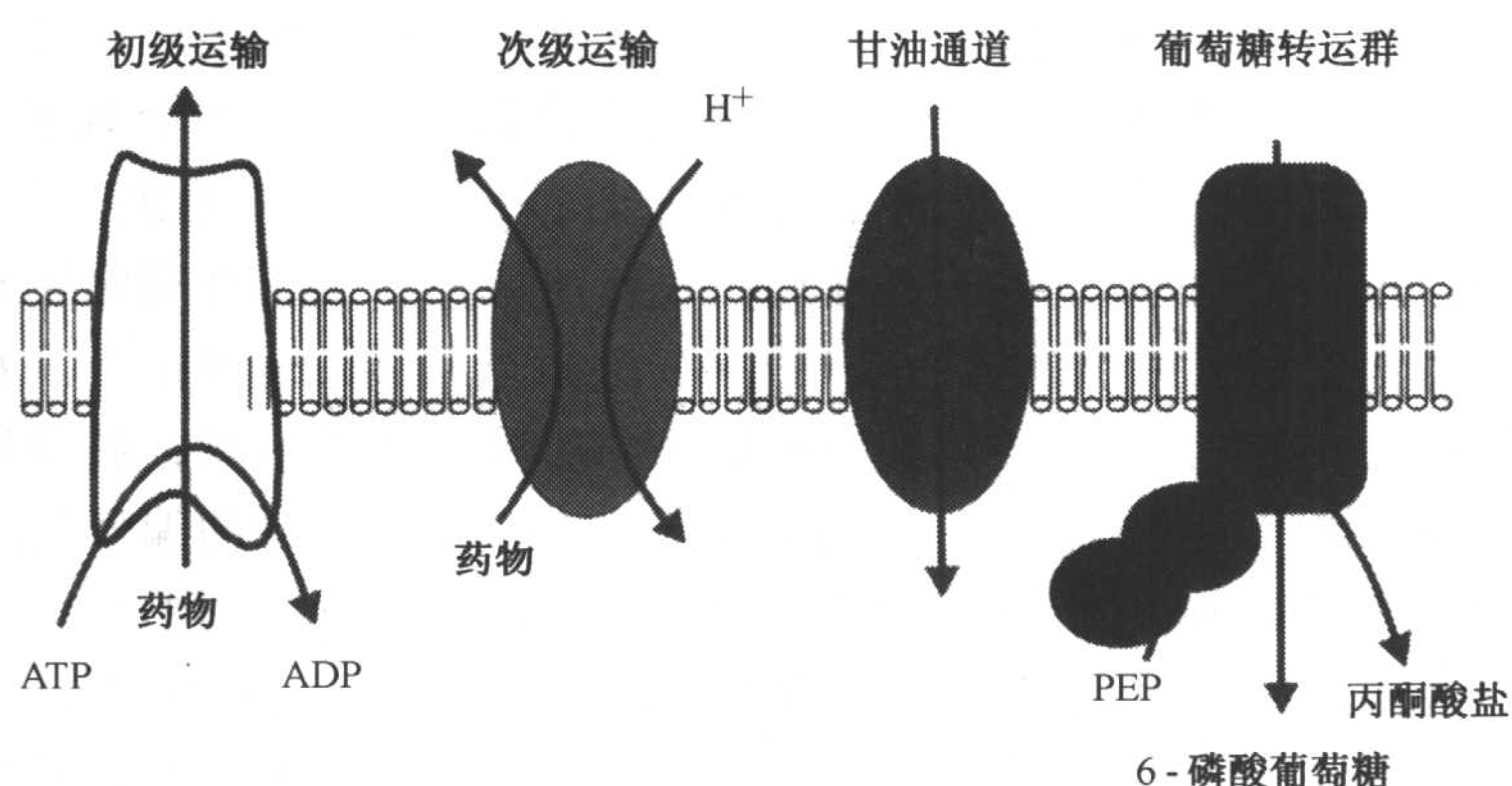


图1　四种主要膜转运蛋白分类的范例。初级膜转运蛋白：乳酸乳球菌（*Lactococcus lactis*）LmrP多药排出泵（multidrug efflux pump）；次级膜转运蛋白：金黄色葡萄球菌（*Staphylococcus aureus*）QacA多药排出转运蛋白（multidrug efflux transporter）；通道：大肠杆菌（*Escherichia coli*）GlpF甘油通道；基团转移：大肠杆菌（*Escherichia coli*）PtsG/Crr葡萄糖PTS转运蛋白。

典型细胞质膜转运系统由至少一个膜定位蛋白（membrane-localized protein）组成，而膜定位蛋白通常是由几个跨膜的α螺旋构成，这种结构组成给研究膜转运系统造成了

麻烦，它们的疏水性和只有在加入表面活性剂的条件下，才显现的可溶性给蛋白的纯化和结晶造成了困难。尽管使用高通量方法在获取结核分枝杆菌（*Mycobacterium tuberculosis*）的动力敏感 MscL 通道[3]，浅青链霉菌（*Streptomyces lividans*）KcsA 钾离子通道[4]，骨骼肌肌浆微管网状组织的 P 型钙 ATP 酶（calcium P-type ATPase）[5]，大肠杆菌 ATP 驱动的脂质翻转酶（lipid flippase）MsbA[6]和质子驱动排出多种药物转运蛋白 AcrB[7]等转运蛋白的研究中取得了显著进展，但有关膜转运蛋白三维结构的数据依然很少。

在用传统实验方法研究膜转运蛋白困难重重的情况下，基因组/生物信息学分析成为很具吸引力的选择。已经测序 140 多种微生物基因组，同时还有 300 多个由公共经费资助的微生物基因组测序计划，正在全世界范围内进行[8]（http://www.tigr.org/tdb/mdb/mdbcomplete.html）。而且，这些基因组学研究，涵盖了分属不同种系发生范围广阔的微生物种群，这就为不同种群微生物生活方式的比较基因组学分析提供了可能。

膜转运蛋白的基因组学分析

此前曾对一系列已完成基因组测序的微生物，进行了系统转运蛋白基因组水平的比较[1,9]，还有针对一些特殊基因组一些类似的研究[10]。针对每个已测序微生物都编制了一个完整的阐述预测膜转运蛋白及其可能底物特性的目录（http://www.membranetransport.org）。这个分类编辑工作还在继续进行，并且还正在开发一个能更便捷查询的相关数据库。

所有已确定的转运蛋白，是根据一套针对膜转运蛋白及其家族转运蛋白的分类系统（Transporter Classification，TC）进行了分类[9,11]（http://www.biology.ucsd.edu/~msaier/transport/）。TC 系统根据转运蛋白的功能及其种系发生学，对所有已知膜转运蛋白进行了分类，每一种被 TC 系统分类的转运蛋白都有一个由 5 个数字或字母组成的编号：V.W.X.Y.Z。V（数字）代表转运蛋白的种类（例如，通道、次级转运蛋白、初级转运蛋白或迁移蛋白组）；W（字母）代表转运蛋白亚类，在初级主动运输转运蛋白中它代表驱动运输所采用的能量；X（数字）代表转运蛋白家族；Y（数字）代表转运蛋白的亚族；Z（数字）代表转运底物或底物的类型。依此规则，特征明确的大肠杆菌 LacY 乳糖透性酶（LacY lactose permease）[12]，在 TC 系统中的编号为2.A.1.5.1，2 代表它是一种次级转运蛋白，A 代表它是一个单向（uniporter）/同向（symporter）/反向转运蛋白（antiporter），1 表明它属主要易化超级家族（major facilitator superfamily，MFS），5 表明它属该家族中一个寡糖同向转运蛋白亚族，最后的 1 表明它是一个乳糖/质子同向转运蛋白。TC 系统中总共描述了 150 个以上的转运蛋白家族。

在已完成基因组测序的微生物中，据预测，细胞膜转运蛋白的数量相差大约 40 倍（图 2A）。将转运蛋白的数量与微生物基因组的大小进行相关性（图 2B）比较可以发现，转运蛋白/Mb（兆碱基）的比例变化范围仅 2～3 倍，这说明微生物中转运蛋白的数量在一定程度上取决于其基因组大小。

对大量微生物类群进行这种比较分析已成为可能，并能从中看出一些趋势。一些土壤/植物中的微生物，如天蓝色链霉菌（*Streptomyces coelicolor*）、铜绿假单胞菌（*Pseu-*

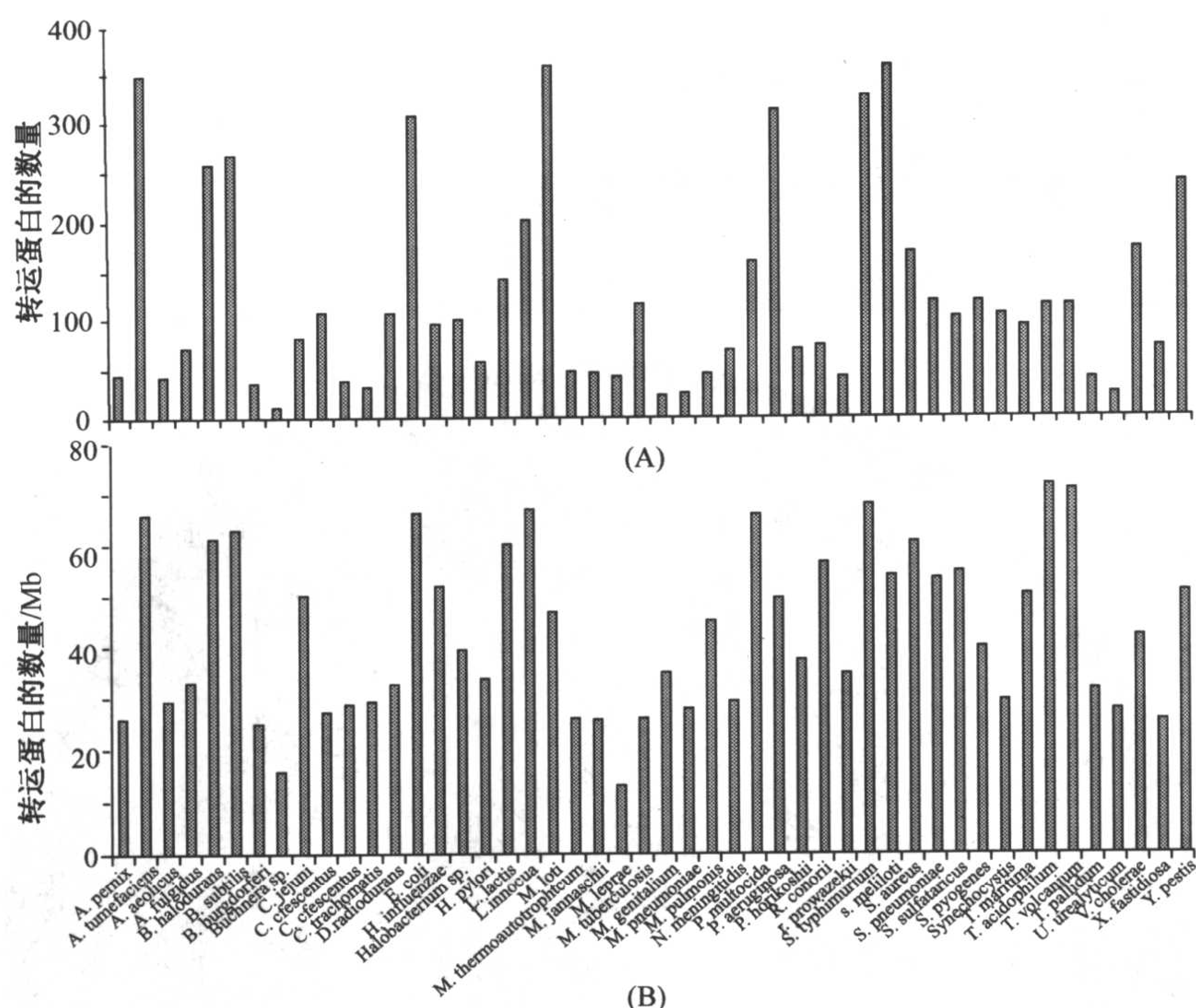


图2　(A) 已测序微生物中预测细胞质膜转运蛋白的数量。(B) 已测序微生物中每兆碱基基因组中预测细胞质膜转运蛋白的数量。

domonas aeruginosa)、百脉根根瘤菌（*Mesorhizobium loti*)、根癌土壤杆菌（*Agrobacterium tumefaciens*）和枯草芽孢杆菌（*Bacillus subtilis*）能够表达相对其基因组大小超量的转运蛋白。一些胃肠道微生物，如大肠杆菌（*E. coli*）和空肠弯杆菌（*Campylobacter jejuni*）也是如此。从绝对数量看，它们中的许多微生物都能产生大量转运蛋白，反映了这些微生物对自然环境中广泛、各异底物的环境适应性，例如，铜绿假单胞菌表现出在不同环境中旺盛的生长能力。

与之相反，大部分胞内病原体或共生生物，如麻风分枝杆菌（*Mycobacterium leprae*)、支原体（*Mycoplasma* spp）和巴克纳氏菌（*Buchnera* sp）都有很低比例的转运蛋白/Mb，这反映了它们生活方式的局限以及相对单一生存环境的本质。大部分测序的古生菌的转运蛋白相对基因组较少，原因在于它们在某些情况下大多倾向于自养而不是异养代谢，有些古生菌则例外，如嗜酸热原体（*Thermoplasma acidophilum*）和火山热原体（*Thermoplasma volcanium*)，相对于基因组大小，它们拥有数量很多的转运蛋白，这反映了它们对外源糖类的需求。

生物体转运蛋白个论

某个生物的全基因组测序，能够在生物信息学预测的基础上重新构建详细的转运和代谢的模型。图3是一个呼吸道病原体肺炎链球菌（*Streptococcus pneumoniae*）的代谢

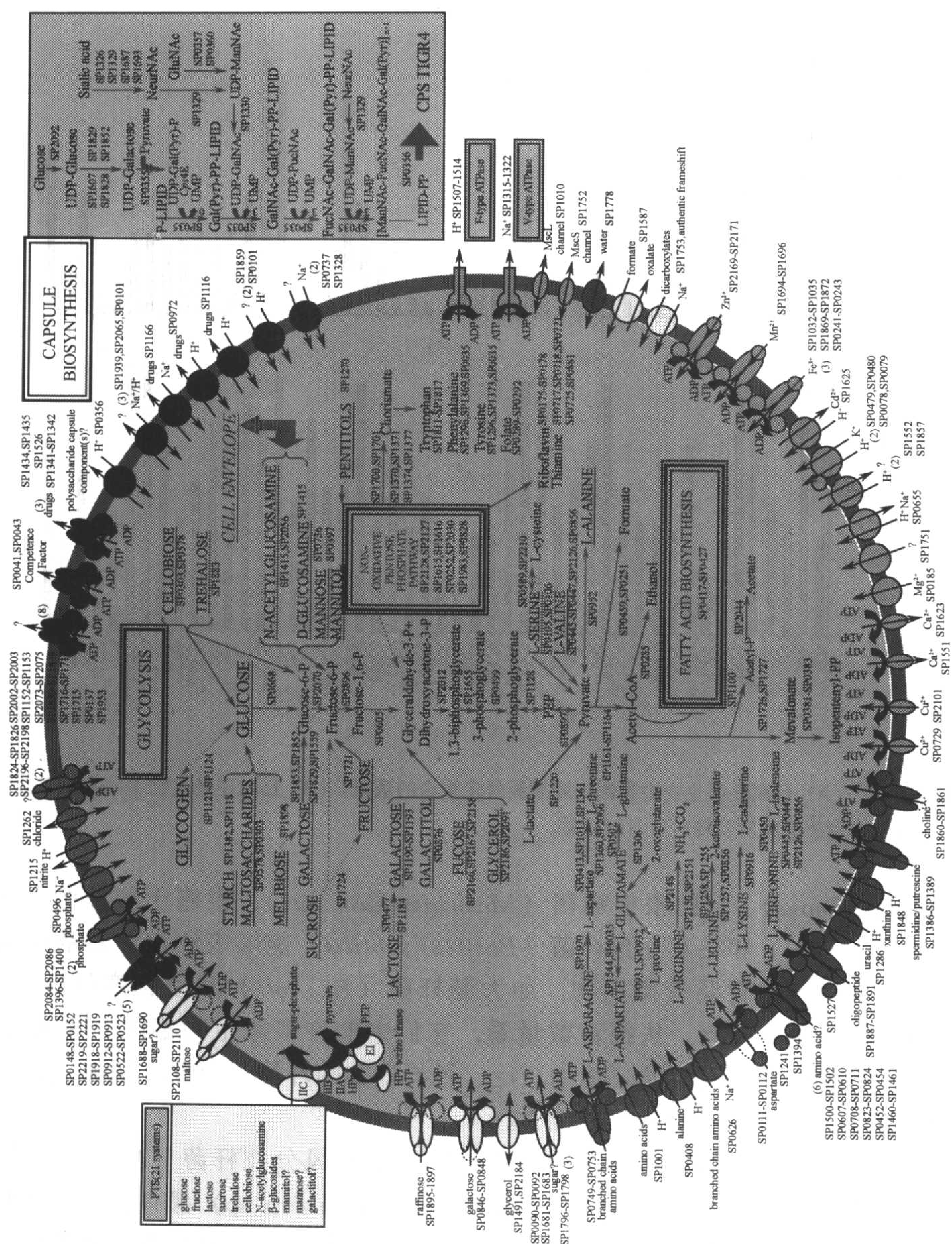


图 3 肺炎链球菌（*Streptococcus pneumoniae*）转运和代谢模型。图中所示为产能，有机化合物的代谢及其荚膜的合成。转运蛋白按照如下所示的底物特异性进行分类：无机阳离子（绿色）、无机阴离子（粉色）、碳水化合物/羧酸盐（黄色）、氨基酸/多肽/胺/嘌呤/嘧啶（红色）、药物排出和其他（黑色）。问号表示所转运的底物不能确定。溶质的输出和输入用穿过转运蛋白的箭头线表示。转运蛋白的能量耦合机制按如下所示：利用蛋白通道溶质的转运用双向箭头线表示；次级转运蛋白用两个箭头线表示，分别代表溶质和伴随转运的离子；ATP 驱动转运蛋白用 ATP 水解反应标明；未知能量耦合机制的转运蛋白用单箭头线标明。组分未知的功能类似多亚基复合体的转运蛋白系统用虚线表示。当具有类似预测底物的多个同源转运蛋白（multiple homologues transporters）存在时，该类型转运蛋白的数目在圆括号内标明。系统的基因编号（SPxxxx）在每个代谢途径或转运蛋白旁边注明；那些用破折号间隔开的代表一系列连续基因（本图再版得到参考文献［13］的授权，版权 2000 美国科学发展协会）。（另见文前彩色插图 7-3）

和转运分析实例[13]。在有关代谢的章节里可以找到真核寄生生物恶性疟原虫（*Plasmodium falciparum*）和细菌恶臭假单胞菌（*Pseudomonas putida*）的转运和代谢的详细模型。

大多数情况下，不能有把握地预测转运蛋白精确的底物特异性（图 3 中，预测的特异性标有问号，或是用一个类似于“氨基酸”一般性预测来代替精确预测），尽管如此，预测的转运蛋白和代谢途径之间有良好的相关性，反映了这类分析能很好地帮助从总体上认识生物机能。

在肺炎链球菌中，预计有 30% 以上的转运蛋白为运输糖类服务，这个比例在已测序生物中非常高[13]。尤其是它利用了 21 个不同 PTS 型糖类转运蛋白，以及多重 ATP 结合区（ATP binding cassette，ABC）超家族糖类转运蛋白。这与其他一些代谢途径的存在相吻合，如直接参与糖酵解途径的多个糖代谢途径和经磷酸戊糖途径的戊糖醇途径。此外，肺炎链球菌还编码一部分胞外酶，如 N 乙酰葡萄糖氨酶（N-acetylglucosaminidase），α 和 β 半乳糖苷酶（α-and β-galactosidase），内切糖苷酶（endoglycosidase）、水解酶（hydrolase）、透明质酸酶（hyaluronidase）和神经氨酸苷酶（neuraminidase），使寄主胞壁质、糖脂和透明质酸等糖类聚合物分解成相应的糖。肺炎链球菌中存在大量的糖及多元醇转运蛋白，或许反映了它们分解寄主糖类聚合物提供营养，以及破坏寄主组织以定殖的能力，肺炎链球菌还可能利用自己的糖类转运蛋白和分解途径来循环利用它自身的多糖荚膜。

其他已测序的呼吸道病原体，如脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）和流感嗜血菌（*Haemophilus influenzae*），有相对较少的糖类转运蛋白，却表现出对羧酸盐/酯和其他碳化合物的偏好[1]。这表明肺炎链球菌生活在一个特殊的呼吸道微环境中，其他链球菌也表现出对糖代谢转运的高度依赖性。

肺炎链球菌也有一部分氨基酸、多胺、嘌呤和嘧啶转运蛋白，目前只确定了两个羧酸盐/酯转运蛋白，其中一个在肺炎链球菌菌株 TIGR4 中有一个移码突变。与大量糖类转运蛋白相比，缺乏羧酸盐/酯转运蛋白和三羧酸（tricarboxylic acid，TCA）循环欠缺与这种兼性厌氧生物的发酵型生活有关。肺炎链球菌的无机离子转运蛋白相对有限，包括 Mn^{2+} 和 Zn^{2+} 转运蛋白，据报道它们与毒力有关[14,15]。有三个铁螯合 ABC 转运蛋白和三个磷酸盐 ABC 转运蛋白可能参与从寄主体内摄取铁和磷酸盐，并与毒力有关。

总览肺炎链球菌的转运和代谢，可提供一个用生物信息分析获得深入、全面的图景实例。须强调的是，对转运和代谢能力交互关联的预测，以及此前从文献中得到的生理信息在很大程度上加深了这种分析（见第 6 章）。

对一种生物转运和代谢的机能想构建定性和定量的计算机模型现在已经可行了。例如，将膜转运蛋白信息合并入大肠杆菌代谢的 EcoCyc 数据库[16,17]（Paulsen，IT and Karp，PD，2003 年未发表数据）。EcoCyc 中有已知及预测的大肠杆菌转运蛋白，以及建立在原始文献基础上对每一个转运蛋白的详细描述。在 EcoCyc 资料库里，根据大肠杆菌转运蛋白和代谢途径的数据结构组建方式，可以对转运和代谢进行复杂的交互查询，也使定性分析这些过程成为可能。

另一个例子是幽门螺旋杆菌（*Helicobacter pylori* 26695）[18]和流感嗜血菌（*H. influenzae*）[19,20]的代谢和转运网络的重新构建。在这个例子里，基数控制模型、极端途

径分析、流量平衡分析都用来分析计算机模型的特征，并预测菌体生长最少底物的需求及预测可能作为主要碳源的底物。计算机模拟（*in silico*）基因敲除和对这些“突变株”在不同培养基中的生长模拟，可以用来预测必需基因，在有些例子中，这些基因的必要性已经用体外基因敲除研究所证实[18]。

底物特异性及生物能学比较

整套预测转运蛋白的底物特异性和生物能学，可以在不同基因组之间进行比较。此前的分析表明，利用转运蛋白能量耦合机制，可以反映出一个生物的整体代谢和生物能学[1,9]，例如，缺乏 TCA 循环和电子传递链的生物，如支原体（*Mycoplasma* spp）、螺旋体（*Spirochetes*）、海栖热袍菌（*Thermotoga maritima*）和肺炎链球菌（*S. pneumoniae*），都高度依赖 ATP 驱动转运蛋白，因为它们通过间接方法来获得质子运动的动力。一些光合作用生物，如集胞蓝细菌（*Synechocystis* PCC6803）、聚球蓝细菌（*Synechococcus* WH8102）和绿硫菌（*Chlorobium tepidum*）也具有绝大多数 ATP 驱动转运蛋白，因为它们有通过光合作用合成 ATP 的能力。

有些例子，偏嗜某一特定类型转运蛋白的基本原理尚不清楚，例如，一组 α 多形菌（α-proteobactia），包括根癌土壤杆菌、百脉根根瘤菌（*M. loti*）、苜蓿根瘤菌（*Sinorhizobium meliloti*）和猪布鲁氏菌（*Brucella suis*）高度依赖 ATP 驱动转运蛋白，可至今仍无确切生物能学原理可以解释其原由，一种可能解释是相对于次级转运蛋白，初级转运蛋白经常具有更高的底物亲和力，所以这些生物对于高亲和力转运有一种特殊需求[21]。

对已测序膜转运蛋白的整体底物特异性分析表明，它们具有高度的异质性差别，这在很大程度上反映出它们在自然生活环境中底物的丰富多样性[1]，例如，已测序的专性胞内寄生生物立克次氏体（*Rickettsia*）和衣原体（*Chlamydia*）具有相当有限的转运能力，几乎不能转运自由糖，但有大量氨基酸和核苷转运蛋白，使它们能从寄主中摄取氨基酸和核苷。与此相反，已经讨论过的肺炎链球菌和其他链球菌，还有海栖热袍菌等生物，它们拥有很大比例的糖类转运蛋白，能够分解并利用大量复杂的植物多糖。此外，还有粪肠球菌（*Enterococcus faecalis*），糖类转运蛋白促进了对胃肠道中没被吸收糖的发酵。

根瘤生物，如根癌土壤杆菌、百脉根瘤菌和苜蓿中华根瘤菌，它们分享大量氨基酸、肽和糖类转运蛋白，这也反映了根际营养丰富的环境。其他土壤/植物中的微生物，如枯草芽孢杆菌，铜绿假单胞菌和恶臭假单胞菌都有非常广泛的转运能力，尽管后两者极端缺乏糖类转运蛋白，但它们能广泛利用包括芳香族底物在内的其他碳化合物，尤其是恶臭假单胞菌。在土壤/植物中的细菌，多药转运蛋白基因普遍存在，并在一些胞内病原体基因组中过量存在，例如伯氏考克斯体（*Coxiella burnetii*）和普氏立克次体（*Rickettsia prowazekii*），以便于它们排出寄主产生的抗菌肽[22]。

相近物种/菌株之间转运蛋白的比较

随着基因组测序的成熟，从不同水平对相近种间或株间的细菌基因组进行比较分析

已成为可能，例如，猪布鲁氏菌（*B. suis*）[23]和山羊布鲁氏菌（*Brucella melitensis*）[23a]基因组全序列已经发表，这两种病原菌都是人畜普鲁氏病的病原[24]，但它们的毒力和寄主偏嗜性不同，山羊布鲁氏菌比猪布鲁氏菌对人类毒性更大，猪布鲁氏菌偏嗜猪作为寄主，而山羊布鲁氏菌则偏嗜山羊和绵羊。

且不论表型上的这些差异，这两种细菌的基因组具有非常高度保守的序列和同线性。在它们3.31Mb的基因组中，有3.2Mb以上的片段完全保守[23]，在每个基因组内都有一小部分特殊区域或“岛”被确认，许多“岛”是噬菌体整合造成的。猪布鲁氏菌（*B. suis*）中发现了42个，而山羊布鲁氏菌（*B. melitensis*）中发现了32个特殊基因，其中大部分的功能尚不清楚[23]。猪布鲁氏菌（*B. suis*）的两个独特岛包含预测的ABC氨基酸转运蛋白基因，其中的一个岛由于残存ABC氨基酸转运蛋白基因簇的片段，而被认为是山羊布鲁氏菌（*B. melitensis*）基因组片段缺失造成的。猪布鲁氏菌（*B. suis*）中两个ABC氨基酸转运蛋白的存在，可以解释猪布鲁氏菌（*B. suis*）利用鸟氨酸、瓜氨酸、精氨酸和赖氨酸的能力，而山羊布鲁氏菌（*B. melitensis*）则没有这些功能，这些差异是否影响寄主偏嗜性和毒力尚待研究。

对流产布鲁氏菌（*Brucella abortus*）（Halling 等未发表）和绵羊布鲁氏菌（*Prucella ovis*）（Paulsen 等未发表）全基因组序列的探索正在进行中。因此，很快就能将这些观察结果推广到布鲁氏菌属的其他种。在一些多个基因组序列已知的种或属中，也能进行类似的分析，如大肠杆菌及相关的肠细菌、各种链球菌和衣原体、结核分枝杆菌的各个菌株等等。可以预见，随着序列数据容量的不断增加，这类在相近菌株/种间进行的比较会越来越有信息价值。

膜转运蛋白的种系发生/种系发育分析

另一种类型的基因组比较分析，是通过种系发育方法来重新构建转运蛋白的家族进化史，用研究基因缺失、水平转移、基因复制和扩增等方法，来详细检测所有已测序基因组中转运蛋白的特定类型或家族。

先前对转运蛋白家族种系发生的研究，揭示了底物特异性是一个高度保守的进化特性，就是说，具有相似底物特异性的转运蛋白有聚集成簇的倾向[1,26~29]，图4中MFS代表成员的种系发生树就是一个实例，MFS的种系发育分析已在该超家族中确定了34个不同的亚族[28,29]，其中15个显示在图4的种系发生树中，MFS已包含上千个确定的成员，与ABC超家族一起成为自然界最大的两个转运蛋白家族之一。

尽管MFS成员都是次级转运蛋白，但它们的功能不同，分为转运糖、羧酸盐/酯、氨基酸、芳香化合物、药物、无机阴阳离子和其他不同化合物的转运蛋白[28]。图4种系发生树中大部分群或亚家族是特异针对一类特定底物（如药物、单糖、核苷等等）。图上有一个家族成员没有任何特征，另一个命名为代谢产物摄取家族显示出相对广泛的底物特异性。因此，种系发育分析为功能预测提供了帮助，对任何新转运蛋白，至少在确定其底物的类别方面有帮助。

用生物信息分析对转运蛋白做精确预测会遇到很多问题。例如，单碱基对的置换既能改变大肠杆菌LacY乳糖转运蛋白的底物特异性，使其转运麦芽糖、阿拉伯糖或其他

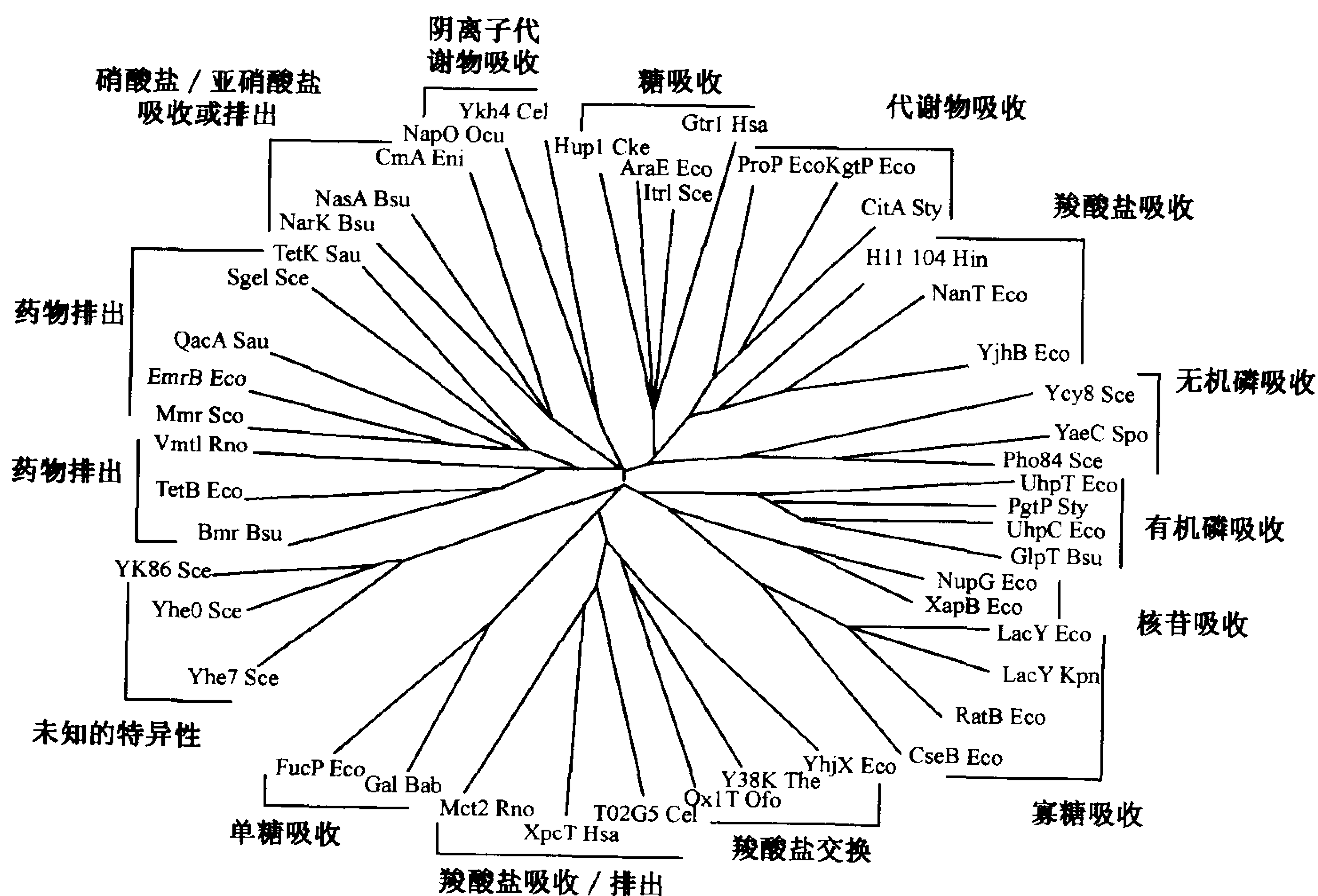


图 4　MFS 代表成员的无根种系发生树。蛋白和菌株缩写同 Pao 等[28]和 Saier et al.[29]。在所有已知 34 个 MFS 亚族中，本发生树中包括了 15 个，并标明了它们已知底物的特异性。

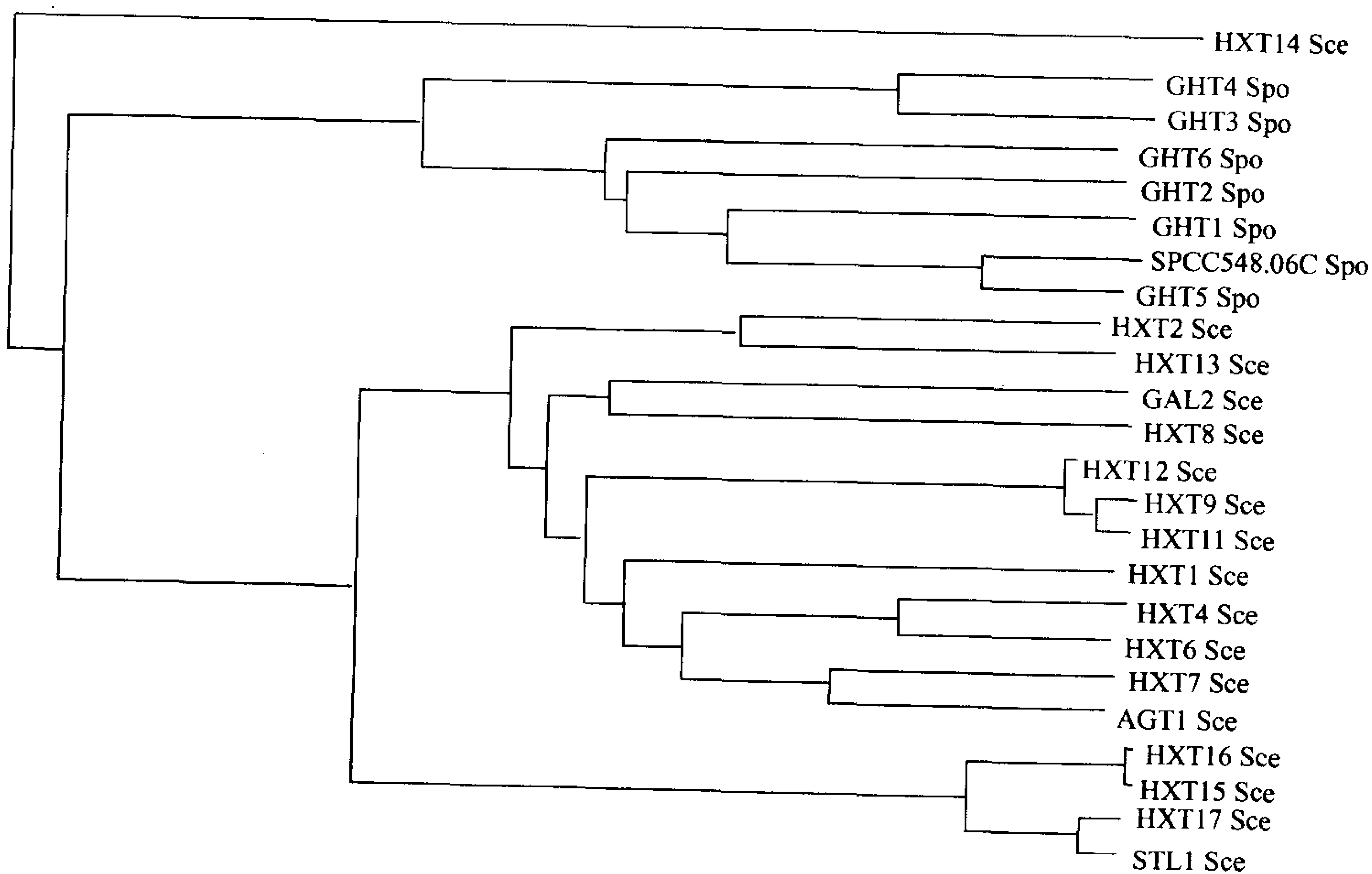


图 5　真菌糖类转运蛋白种系发生树中的部分分支（Greenberg and Paulsen，未发表数据）。图中显示了酿酒酵母的糖转运蛋白（Sce）和裂殖酵母的糖转运蛋白（Spo）。

糖[30~33]，又能改变金黄色葡萄球菌（*Staphylococcus aureus*）QacA[34]和枯草芽孢杆菌 Bmr[35]多药转运蛋白的药物特异性，这表明对同类化合物，转运蛋白较容易调整其底物特异性。但是，对 LacY 的小小改动却不能把它变为多药转运蛋白，从进化时间衡量，转运蛋白进化成转运完全不同另一类底物的概率不高。对单个 MFS 亚族种系发生的分析表明，在至少 3 种情况下，糖类的转运能力起源并维持在 3 种不同进化阶段，而转运其他类型底物，如核苷或无机磷酸盐的能力，仅起源并维持在一个进化阶段。

已测序基因组中特定转运蛋白家族进行详细种系发育分析，对功能预测很有价值。图 5 表明了酿酒酵母（*Saccharomyces serevisiae*）和裂殖酵母（*Schizosaccharomyces pombe*）的已知和预测糖类转运蛋白的部分种系发生树，可清楚看出，大部分酿酒酵母 HXT 转运蛋白和裂殖酵母 GHT 转运蛋白的发生，都是由这两种酵母在进化史上分离后而独立演化产生的。图 5 中显示的这些酵母转运蛋白彼此不是直系，由此推测这两种生物的糖类转运蛋白之间在进行精确功能信息分析时须持谨慎态度。

膜转运蛋白的功能基因组学分析

大量基因组序列数据使功能基因组学分析，如微阵列表达分析、大规模基因敲除或表达研究等，成为研究生物学问题越来越有吸引力的手段。目前，仅在小规模范围内，对不同微生物的特定转运蛋白家族，或类型系统性的基因敲除或过量表达进行了研究。

用这种方法的一个实例，是对粪肠球菌（*E. faecalis*）[36]的 30 个推测多药转运蛋白的系统性插入失活，然后测试这 30 个突变株对 28 种不同抗微生物药物或化合物的敏感性。其中 4 个 ABC 转运蛋白基因的破坏，导致细菌对至少一种抗微生物药物的敏感性显著增加，从而确定它们可能为编码药物排出的转运蛋白。一个对酿酒酵母的类似分析，将酵母 ABC 药物转运蛋白和调节因子进行单个及多个删除，然后用一组抗真菌化合物来对突变株进行筛选[37]。结果表明，许多这类转运蛋白有重叠的特异性，而当多个基因被删除时会产生对药物超敏感的突变株。

同源和异源的表达还用于研究新的转运机制，例如 Nishino 和 Yamaguchi[38]，从大肠杆菌中克隆了 37 个预测用于药物排出的转运蛋白，并利用一株大肠杆菌药物敏感突变株，对它们的抗药性进行筛选。在 37 个推测的药物排出转运蛋白中，有 20 个加强了细菌对一种或多种抗生素的抗性。在这 20 种药物转运蛋白中，包括 7 种新药物转运蛋白和 6 种底物特异性被扩大的转运蛋白。与此相仿，大肠杆菌中的异源表达，还用于发现一系列病原菌的多重 SMR 家族的多药转运基因[39]。

小规模转运蛋白功能基因组学分析并不仅仅聚焦于药物转运蛋白，例如，系统性插入失活，用于研究集胞蓝细菌（*Synechocystis* PCC6803）[40]预测氨基酸转运蛋白基因。我们正在综合利用异源表达、基因敲除和微阵列分析等方法来研究海洋聚球蓝细菌（*Synechococcus*）WH8102 转运蛋白的组成（Palenik，Brahamsha，and Paulsen，未发表资料）。

微阵列分析开始为转运蛋白功能和调节的研究提供大量数据，这里有用大肠杆菌微阵列研究表达的几个例子。对大肠杆菌的 MarA 全局调节因子组成型表达微阵列分析，确定了除多药排出基因 *acr*AB 和 *tol*C 之外的另外两个基因表达。受 MarA 影响的流出

系统 YadGH 和 YdeA[41]。对大肠杆菌双组分信号传递基因突变的微阵列表达分析，发现了大量受双组分系统调控的大肠杆菌转运蛋白[42]。对野生型大肠杆菌及一个敲除亮氨酸应答调节蛋白（在调节稳定期基因表达中起着重要作用）的突变株进行基因表达分析，发现了一些在稳定期受亮氨酸应答调节蛋白正调节的转运蛋白，包括渗透压保护物质（osmoprotectant）转运蛋白[43]。对 *evg*A 应答调节基因的敲除突变株或过量表达该基因的大肠杆菌菌株的微阵列研究表明，*evg*A 对许多抗酸基因和药物排出泵基因 *emr*K 和 *yhi*UV 起调节作用[44]。微阵列表达分析对其他细菌，如枯草芽孢杆菌（*Bacillus subtilis*）[45]、结核分枝杆菌（*M. tuberculosis*）[46]、希瓦氏菌（*Shewanella oneidensis*）[47]和菊欧文氏菌（*Erwinia chrysanthemi*）[48]转运蛋白的调节有了更深入的了解。

大规模基因敲除研究，开始使膜转运蛋白基因的功能和必要性研究有了起色。生殖道支原体（*Mycoplasma genitalium*）和肺炎支原体（*Mycoplasma pneumoniae*）的基因组很小，用转座子插入突变来确定它们生长所必需的最少基因，结果表明生殖道支原体只有 265～330 个必需的基因[49]，并得到了几个转运蛋白基因敲除的突变株，包括预测的 ABC 药物排出泵和 PTS 果糖吸收系统。

通过敲除酵母 96％预测 ORF 以及将那些独特的脱氧核糖核酸序列作为“分子条形码（molecular bar codes）标记”[50]，构建了大量的酿酒酵母突变株。对在丰富培养基中表现出缓慢生长表型的 15％纯合子突变株的检测，发现了 F－和 V－型 ATP 酶亚单位，以及其他一些转运蛋白。此外，还发现了酵母在不同环境条件下，最佳生长所必需的其他许多转运蛋白。

结论

基因组时代的一大挑战是充分利用从基因组测序、微阵列实验和高通量功能基因组学所得到的大量数据。在膜转运方面，有两个平行的目标，一是了解生物体中所有转运蛋白，另一是对生物的转运和代谢建立计算机模型，这两个目标已越来越接近。近来，在膜转运蛋白结构分析方面的研究成果，使其在分子水平上对底物的识别和膜转运系统的运输有了深入了解。然而，还有许多对膜转运系统进行全面了解的研究尚未完成，即使在大肠杆菌中，几乎有一半预测的膜转运系统还需要通过实验来验证其特征。

（江　昊　译）

参考文献

1. Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH Jr. Microbial genome analyses: comparative transport capabilities in 18 prokaryotes. J Mol Biol 2000; 301:75–100.
2. Saier MH Jr. Vectorial metabolism and the evolution of transport systems. J Bacteriol 2000; 182:5029–5035.
3. Chang G, Spencer RH, Lee AT, Barclay MT, Rees DC. Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. Science 1998; 282: 2220–2226.

4. Doyle DA, Morais Cabral J, Pfuetzner RA, et al. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 1998; 280:69–77.
5. Toyoshima C, Nakasako M, Nomura H, Ogawa H, et al. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. Nature 2000; 405:647–655.
6. Chang G, Roth CB. Structure of MsbA from *E. coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. Science 2001; 293:1793–1800.
7. Murakami S, Nakashima R, Yamashita E, Yamaguchi A. Crystal structure of bacterial multidrug efflux transporter AcrB. Nature 2002; 419:587–593.
8. Nelson KE, Paulsen IT, Heidelberg JF, Fraser CM. Status of genome projects for nonpathogenic bacteria and archaea. Nat Biotechnol 2000; 18:1049–1054.
9. Paulsen IT, Sliwinski MK, Saier MH Jr. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J Mol Biol 1998; 277:573–592.
10. Meidanis J, Braga MD, Verjovski-Almeida S. Whole-genome analysis of transporters in the plant pathogen *Xylella fastidiosa*. Microbiol Mol Biol Rev 2002; 66:272–299.
11. Saier MH Jr. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 2000; 64:354–411.
12. Kaback HR, Sahin-Toth M, Weinglass AB. The kamikaze approach to membrane transport. Nat Rev Mol Cell Biol 2001; 2:610–620.
13. Tettelin H, Nelson KE, Paulsen IT. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 2001; 293:498–506.
14. Jakubovics NS, Smith AW, Jenkinson HF. Expression of the virulence-related Sca (Mn2+) permease in *Streptococcus gordonii* is regulated by a diphtheria toxin metallorepressor-like protein ScaR. Mol Microbiol 2000; 38:140–153.
15. Dintilhac A, Alloing G, Granadel C, Claverys JP. Competence and virulence of *Streptococcus pneumoniae*: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. Mol Microbiol 1997; 25:727–739.
16. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. The EcoCyc and Meta Cyc databases. Nucleic Acids Res 2000; 28:56–59.
17. Karp PD, Riley M, Saier M. The EcoCyc Database. Nucleic Acids Res 2002; 30:56–58.
18. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO. Genome-scale metabolic model of *Helicobacter pylori* 26695. J Bacteriol 2002; 184:4582–4593.
19. Edwards JS, Palsson BO. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J Biol Chem 1999; 274:17,410–17,416.
20. Schilling CH, Palsson BO. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. J Theor Biol 2000; 203:249–283.
21. Wood DW, Setubal JC, Kaul R, et al. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science 2001; 294:2317–2323.
22. Paulsen IT, Lewis K. Microbial multidrug efflux: introduction. J Mol Microbiol Biotechnol 2001; 3:143–144.
23. Paulsen IT, Seshadri R, Nelson KE, et al. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci USA 2002; 99: 13,148–13,153.
23a. Del Vecchio VG, Kapatral V, Redkar RJ, et al. The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. Proc Natl Acad Sci USA 2002; 99:443–448
24. Smith LD, Ficht TA. Pathogenesis of *Brucella*. Crit Rev Microbiol 1990; 17:209–230.
25. Eisen JA, Hanawalt PC. A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res 1999; 435:171–213.

26. Jack DL, Paulsen IT, Saier MH. The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. Microbiology 2000; 146 (Pt 8):1797–1814.
27. Paulsen IT, Brown MH, Skurray RA. Proton-dependent multidrug efflux systems. Microbiol Rev 1996; 60:575–608.
28. Pao SS, Paulsen IT, Saier MH Jr. Major facilitator superfamily. Microbiol Mol Biol Rev 1998; 62:1–34.
29. Saier MH Jr, Eng BH, Fard S, et al. Phylogenetic characterization of novel transport protein families revealed by genome analyses. Biochim Biophys Acta 1999; 1422:1–56.
30. King SC, Wilson TH. Characterization of *Escherichia coli* lactose carrier mutants that transport protons without a cosubstrate. Probes for the energy barrier to uncoupled transport. J Biol Chem 1990; 265:9645–9651.
31. Goswitz VC, Brooker RJ. Isolation of lactose permease mutants which recognize arabinose. Membr Biochem 1993; 10:61–70.
32. Varela MF, Brooker RJ, Wilson TH. Lactose carrier mutants of *Escherichia coli* with changes in sugar recognition (lactose vs melibiose). J Bacteriol 1997; 179:5570–5573.
33. King SC, Wilson TH. Identification of valine 177 as a mutation altering specificity for transport of sugars by the *Escherichia coli* lactose carrier. Enhanced specificity for sucrose and maltose. J Biol Chem 1990; 265:9638–9644.
34. Paulsen IT, Brown MH, Littlejohn TG, Mitchell BA, Skurray KA. Multidrug resistance proteins QacA and QacB from *Staphylococcus aureus*: membrane topology and identification of residues involved in substrate specificity. Proc Natl Acad Sci USA 1996; 93:3630–3635.
35. Klyachko KA, Schuldiner S, Neyfakh AA. Mutations affecting substrate specificity of the *Bacillus subtilis* multidrug transporter Bmr. J Bacteriol 1997; 179:2189–2193.
36. Davis DR, McAlpine JB, Pazoles CJ, et al. *Enterococcus faecalis* multi-drug resistance transporters: application for antibiotic discovery. J Mol Microbiol Biotechnol 2001; 3:179–184.
37. Rogers B, Decottignies A, Koloczkowski M, Carvajal E, Balzi E, Goffeau A. The pleitropic drug ABC transporters from *Saccharomyces cerevisiae*. J Mol Microbiol Biotechnol 2001; 3: 207–214.
38. Nishino K, Yamaguchi A. Analysis of a complete library of putative drug transporter genes in *Escherichia coli*. J Bacteriol 2001; 183:5803–5812.
39. Ninio S, Rotem D, Schuldiner S. Functional analysis of novel multidrug transporters from human pathogens. J Biol Chem 2001; 276:48,250–48,256.
40. Quintero MJ, Montesinos ML, Herrero A, Flores E. Identification of genes encoding amino acid permeases by inactivation of selected ORFs from the *Synechocystis* genomic sequence. Genome Res 2001; 11:2034–2040.
41. Barbosa TM, Levy SB. Differential expression of over 60 chromosomal genes in *Escherichia coli* by constitutive expression of MarA. J Bacteriol 2000; 182:3467–3474.
42. Oshima T, Arba H, Masuda Y. Transcriptome analysis of all two-component regulatory system mutants of *Escherichia coli* K-12. Mol Microbiol 2002; 46:281–291.
43. Tani TH, Khodursky A, Blumenthal RM, Brown PO, Mathews RG. Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. Proc Natl Acad Sci USA 2002; 99: 13,471–13,476.
44. Masuda N, Church GM. *Escherichia coli* gene expression responsive to levels of the response regulator EvgA. J Bacteriol 2002; 184:6225–6234.
45. Britton RA, Eichenberger P, Gonzalez-Pastor JE, et al. Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of *Bacillus subtilis*. J Bacteriol 2002; 184:4881–4890.
46. Wilson M, DeRisi J, Kristensen HH, et al. Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. Proc Natl Acad Sci USA 1999; 96:12,833–12,838.

47. Beliaev AS, Thompson DK, Fields MW, et al. Microarray transcription profiling of a *Shewanella oneidensis* etrA mutant. J Bacteriol 2002; 184:4612–4616.
48. Okinaka Y, Yans CH, Perra NJ, Keen NT. Microarray profiling of *Erwinia chrysanthemi* 3937 genes that are regulated during plant infection. Mol Plant Microbe Interact 2002; 15:619–629.
49. Hutchison CA, Peterson SN, Gill SR, et al. Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 1999; 286:2165–2169.
50. Giaever G, Chu AM, Ni L, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 2002; 418:387–391.

8
用基因组学分析细菌细胞周期

Michael T. Laub, Harley H. McAdams and Lucy Shapiro

引言

全基因组测序以及对基因组进行整体分析（global analysis）的技术正在使微生物学发生革命性的变化。在研究单个基因或蛋白质时某些无法或难以鉴别的规律、现象和机制可以用整体分析的方法来解决。本章的主题是关于细菌细胞周期（cell cycle）的调控机制，在基因组学与分子遗传学、生物化学和细胞生物学的结合下这一领域正飞速发展。细胞周期在时间水平上的分子调控机制已经在很多真核生物系统中得到了广泛的关注，但是，在原核生物中由于缺乏合适的模式系统（model system）却所知甚少。

然而，革兰氏阴性细菌新月柄杆菌（*Caulobacter crescentus*）被证明是研究此类问题非常适合的一种模式细菌，不仅它的遗传学和细胞生物学已经研究得很清楚，而且更重要的是，获得它的同步化（synchronized）细胞群体很容易。用同步化细胞做实验，可以精确地分析细胞周期在各个时间段发生的变化。新月柄杆菌的全基因组测序[1]已经完成，这为用基因组学方法研究细菌细胞周期打开了方便之门。最近用DNA微阵列（DNA microarray）进行的研究更加深了对细菌遗传网络（genetic network）的理解，并推动了对细菌细胞周期发展分子机制的深入研究。本章就此进行讨论。

新月柄杆菌：研究细菌细胞周期的模式系统

新月柄杆菌生命周期的一步步进行（图1），有一系列形态、代谢和调控变化在精准有序地发生。新月柄杆菌分裂成两个截然不同的子代细胞：一个有运动性的“游动（swarmer）”细胞和一个固着的“柄（stalked）”细胞。处于G1期的细胞有一根极生鞭毛和一些极生纤毛。这些运动的游动细胞不能启动它们环状单染色体的复制。在一些未知信号的诱导下，游动细胞蜕去鞭毛，收回纤毛，并在细胞的同一端长出一根柄，从而分化成柄细胞。柄是细胞囊（cell envelope）的管状延伸物，它的末端有一个似“抓手（holdfast）”的结构，使得细胞能够黏附在多种物质的表面。脱氧核糖核酸（deoxyribonucleic acid，DNA）复制的开始阶段正好与游动细胞到柄细胞的分化阶段相吻合，所以这一阶段可被作为从G1期到S期的转变时期。处于该转变时期的细胞和由母细胞刚刚分裂而产生的柄细胞处于同一细胞周期阶段。柄细胞在S期中，在一定机制的作用下，距离柄的那一端有0.6个细胞长度的地方开始构建FtsZ环（FtsZ ring）[2]从而形成一个不对称的分裂位点，并开始进行细胞质分裂（cytokinesis）。接着，中间凹陷的分裂前母细胞（predivisional cell）开始与柄相对的另一端长出一根新鞭毛。在染色体复制和分配完成后，细胞就完成了分裂并产生形态和作用完全不同的两个子细胞。

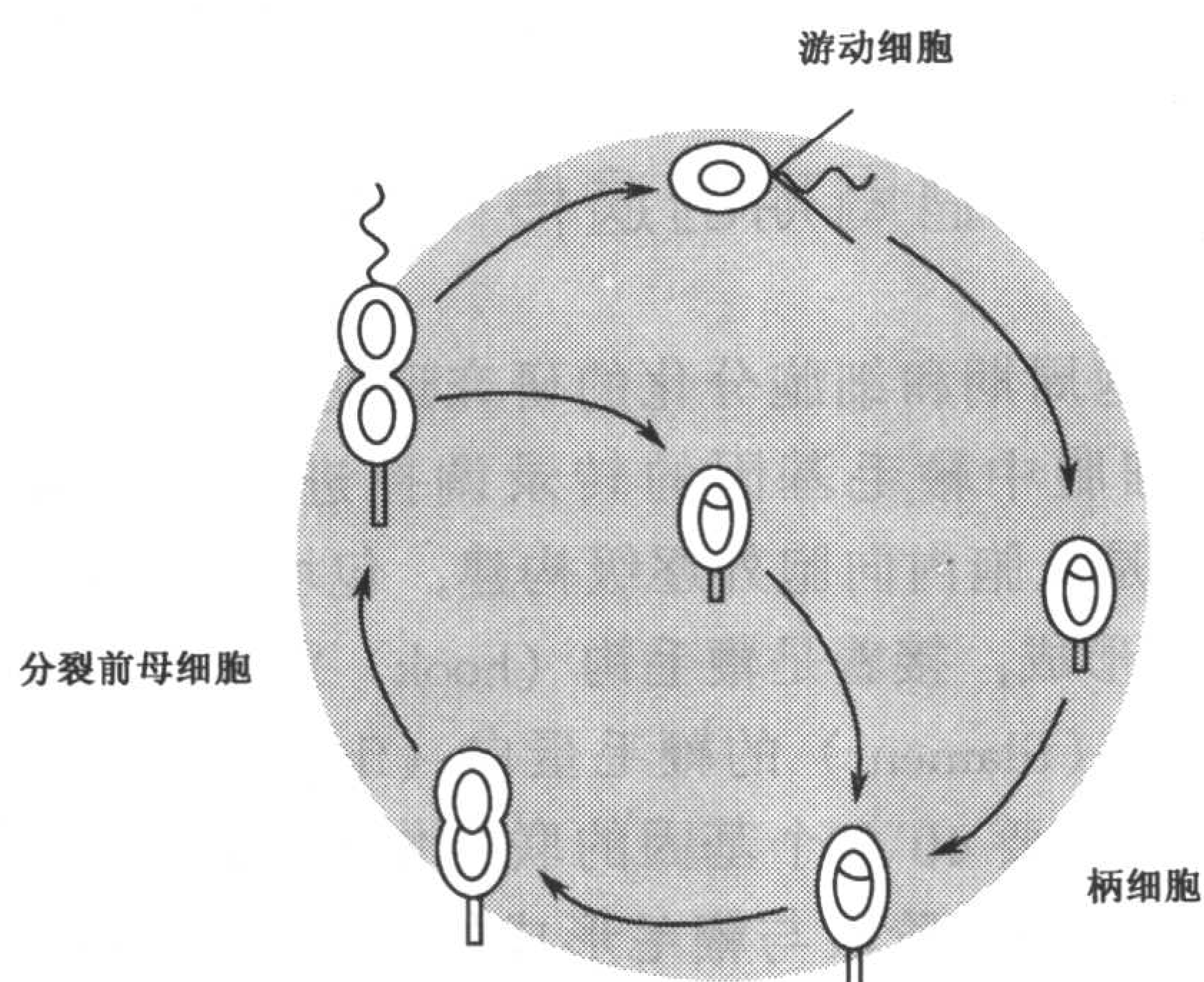


图1 柄杆菌细胞周期的发育和两种不同子代细胞的生成。细胞类型（游动细胞、柄细胞、分裂前母细胞）标注在相应细胞周期阶段的外围。细胞内的环状或 θ 状结构代表指定细胞类型的 DNA 复制阶段

这种不对称分裂的特性可以用于获取同步化细胞，用密度离心（density centrifugation）的方法把小游动细胞与浮力较大的柄细胞和分裂前母细胞分开，然后让这些高纯度的游动细胞群同步进入细胞周期。即使在生长最快的时期［大约 90 分钟的代时 (generation time)］，柄杆菌的染色体在每个细胞周期也只复制一次，并不像大肠杆菌 (*Escherichia coli*）及其他许多细菌那样有重叠复制的现象。总之，柄杆菌细胞易于同步化，并有明显的 G1 期和 S 期，这些特性使柄杆菌成为研究细菌细胞周期的有力工具。

柄杆菌基因组的测序

新月柄杆菌 4.0 Mb 染色体的全序列已经测定[1]。尽管基因组注释加深了对多形杆菌门（Proteobacteria）中整个 α 多形杆菌纲的认识，但真正的挑战还只是从中认识到细菌细胞周期的调控机制和细胞不对称性的建立机制。

搜索同源序列可以预测一些参与特殊细胞周期变化的基因，这类分析尤其适合于那些参与已经研究很深入的一些细胞活动（如 DNA 复制和染色体分离）基因。但是，知道基因组中缺少哪些基因往往与已知基因组中有哪些基因同样重要，也许有时还更重要。例如，在大肠杆菌和枯草芽孢杆菌（*Bacillus subtilis*）中，*minCD* 基因及它的同源基因在选择细胞分裂位点时起关键作用[3~6]。但是，新月柄杆菌的序列里却没有类似的 *minCD* 基因，这就说明它依靠另外一套基因，采用另外一种机制来决定细胞不对称分裂的位点。

尽管基因组序列在这类比较和分析中提供了很多信息，但是仅仅靠序列分析并不能搞清楚细胞周期的分子调控机制。不过，有了全基因组序列，就可以利用很多基因组学新工具来研究推动细胞周期正常运行的调控网络（regulatory network），这些工具中最有用的就是 DNA 微阵列。柄杆菌的全基因组 DNA 微阵列使我们第一次对控制细菌细

胞周期的转录网络（transcriptional network）进行了整体性分析。

用 DNA 微阵列绘制细胞周期的遗传网络图

早年关于柄杆菌细胞周期和细胞分化的研究证明，它的转录调控有时间性[7]，接着对柄杆菌分裂前母细胞中鞭毛基因的转录调控进行了深入研究（综述见参考文献［8］)。它的极生鞭毛是从胞内向胞外逐级构建，包埋在细胞膜中的马达（motor）和基体（basal body）最早形成，接着是鞭毛钩（hook）亚基的外运（export）和组装（assembly），最后是鞭毛丝（filament）的鞭毛蛋白（flagellin）亚基的外运和聚合（polymerization)。鞭毛的形成需要 40 多个基因的联合作用，这些基因组成一个分四阶段的转录梯队（hierarchy），这四个阶段与鞭毛形成的先后顺序相对应。基体的基因（第 2 组）最先转录和表达，如果基因产物只在鞭毛形成后期需要，则最后表达那些基因，例如鞭毛钩（第 3 组）和鞭毛丝（第 4 组）蛋白。另外，每组基因中都有编码反式调控因子的基因，这些反式调控因子能够激活下一组基因的表达，这样就把鞭毛组装的顺序与基因表达的顺序耦联（couple）在一起。

这些前期研究表明，柄杆菌通过转录调控来控制细胞周期的进行，但是，这些研究都是一个基因一个基因地用实验完成的，费时又费力。柄杆菌全基因组序列和 DNA 微阵列的面世，第一次通过信使核糖核酸（messenger ribonucleic acid，mRNA）来深入研究细胞周期依赖性（cell cycle-dependent）基因的表达[9]。在首次整体水平研究中，首先分离了大量处于 G1 期的游动细胞，然后让它们同步进入 150 分钟的细胞周期，每 15 分钟收集一次 RNA 样品，然后再通过 DNA 微阵列把这些 RNA 样品和一个共同的参考样品相比较，就可以得到 2966 个基因在细胞周期中的表达谱（expression profile)。用修饰的 Fourier 技术对这些表达谱数据进行计算分析后发现，约 550 种 mRNA 发生了周期性变化，或者说是细胞周期依赖性变化。这些表达量随时间变化的基因又可归为两种方式：一种是按照表达时间来划分（图 2），其结果表明，在细胞周期中，尽管基因表达时间不同，但它们的表达都是连续进行的，而不是离散的。另一种是按照基因功能和它们在细胞中可能起的作用来归类。

图 3A 展示了一些已知或预测与鞭毛形成有关基因的表达谱。概括地讲，这些基因都是在细胞周期的后期表达，也就是前分裂时期（predivisional stage)，此时正是细胞合成新鞭毛的时候。然而，这些表达谱数据又可以足够精确地显示出这些基因已知的转录顺序，从而证实了鞭毛基因的表达与鞭毛组装之间的共线性关系（colinearity）（图 3A)。这些表达谱可按时间顺序分为四类，正好与先前的四组鞭毛基因的划分吻合（图 3A)[9]。在大肠杆菌中，有人用完全不同的方法证实了类似鞭毛基因在时间上的转录梯队[10]。他们把 14 个鞭毛基因的启动子分别与报告基因 *gfp*［编码绿色荧光蛋白（green fluorescent protein，GFP)］融合，测量 GFP 荧光就可以反映出鞭毛基因的表达。他们发现，这些鞭毛基因的诱导顺序与鞭毛蛋白的组装顺序一致，这些都与在柄杆菌中所观察到的结果相吻合。

在柄杆菌中，与纤毛生成和细胞分裂有关的基因也呈现梯队表达的趋势，表明这些多蛋白结构的组装也依赖相关基因依次有序的表达。已经知道柄杆菌极生鞭毛的生成至

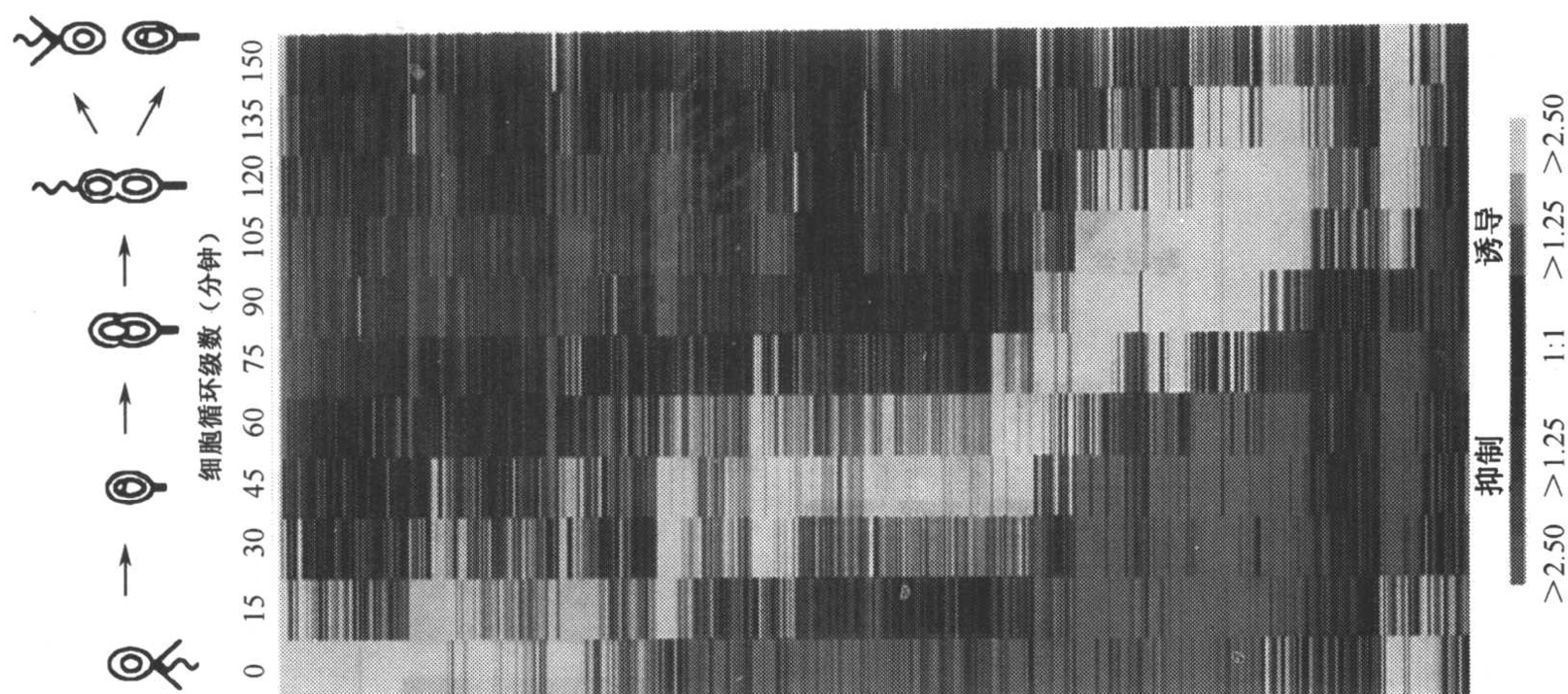


图 2 受细胞周期调控的基因表达。发现的 553 个受细胞周期调控的柄杆菌基因的表达谱以颜色表示。图下方的颜色设计代表 RNA 的相对水平。每个基因的表达谱都是从左到右进行。图上方是细胞周期的示意图以及以分钟数表示的细胞周期进展的时间。（根据许可，依照参考文献［9］重印。American Association for the Advancement of Science 2000 版权所有）（另见文前彩色插图 8-2）

少需要染色体上相邻三个操纵子的表达[11]，这些基因在时间上的依次表达（图 3B）表明，跟鞭毛基因一样，纤毛基因的先后表达顺序反映了在前分裂阶段晚期纤毛形成的过程中，这些基因产物被组装和被利用的先后顺序。柄杆菌的纤毛与第四类型纤毛相关联（type IV-related），纤毛结构中被预测包埋在细胞膜内的部分首先被诱导，并在进入细胞周期后 100 分钟左右达到顶点。再延续 15～30 分钟，原纤毛蛋白酶（prepilin peptidase）的表达也达到了最高值，最后才是纤毛蛋白亚基自身表达的高潮。这些表达谱表明，在柄杆菌中纤毛的生成依照如下的时间调控模式：纤毛蛋白定位亚基（pilin-anchoring subunit）首先表达，以便在细胞一端的细胞囊中建立纤毛组装的位点，接着是原纤毛蛋白酶 CpaA 的表达，有它才能对原纤毛蛋白进行加工。原纤毛蛋白亚基基因 *pilA* 最后表达，*pilA* 被 CpaA 多肽酶切割，生成纤毛蛋白的单体，单体进一步聚合成锚在细胞膜上的纤毛。

与此类似，细胞分裂有关的基因（5 个被发现受细胞周期的调控：*ftsZ*，*ftsI*，*ftsW*，*ftsQ* 和 *ftsA*）也是依次表达的，再一次证明时间上的依次转录对亚细胞结构的正确组装起着至关重要的作用。以下证据也支持这一假定：FtsZ 在新生细胞分裂位点的定位和组装完成后，FtsQ 和 FtsA 才能在细胞质分裂环中组装并具有活性[12,13]，与此一致，*ftsZ* 在柄细胞中达到表达高峰，比 *ftsQ* 和 *ftsA* 的表达（前分裂细胞晚期）早得多。

按照基因功能将表达量随时间变化的基因归类后发现，除了结构基因和表现型基因外，还有很多参与其他细胞活动的基因在一定程度上也受时间因子的调控，从而表现出阶段性表达谱。值得注意的是，几乎所有 DNA 复制的必需基因都是受细胞周期控制的，与 DNA 复制起始有关的基因在游动细胞中的表达量最高，可能是为了保证在 G1—S 的转型期有足够这些基因的产物。DNA 复制延伸、核苷酸合成和 DNA 修复所必需的

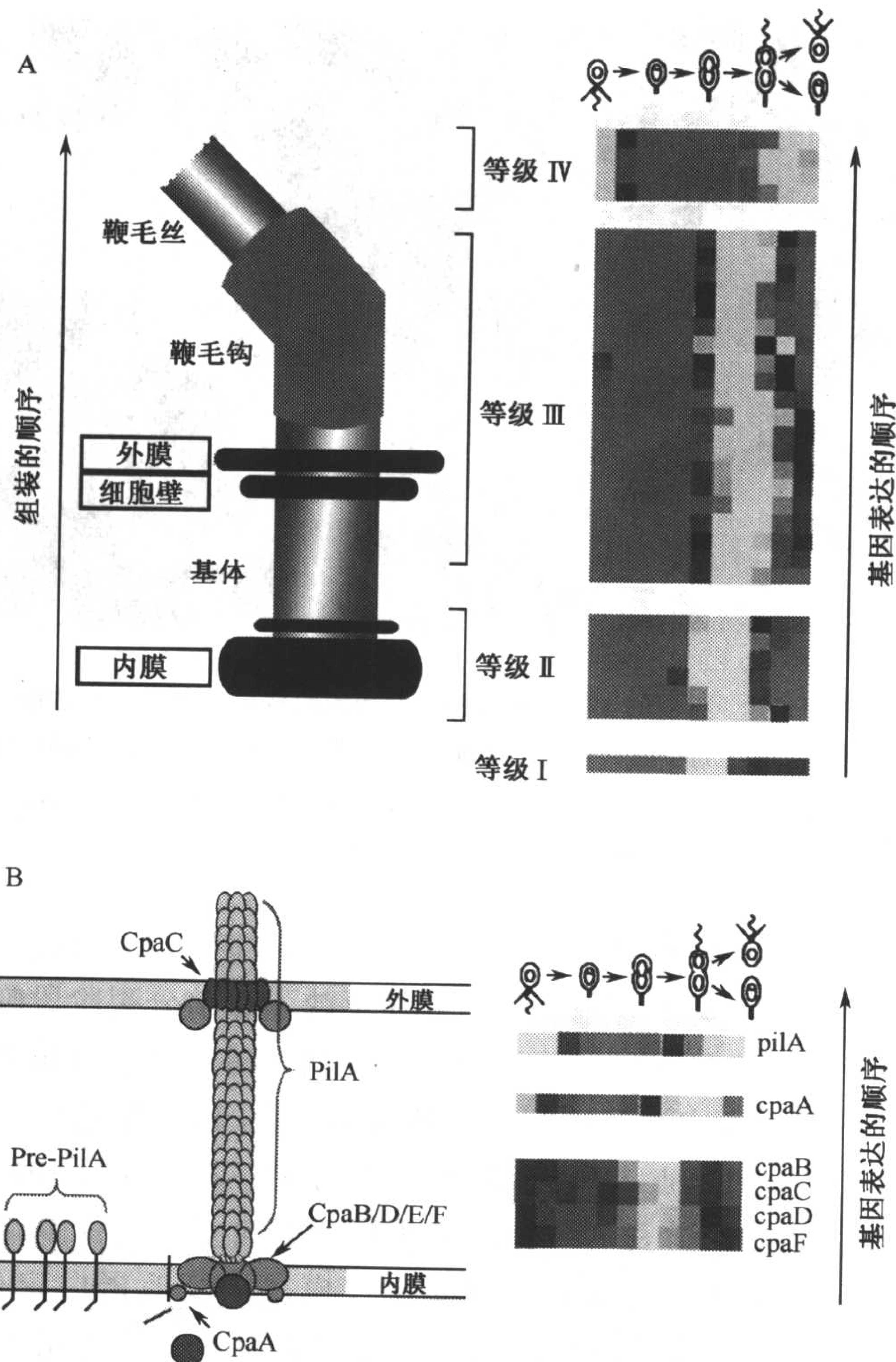


图 3　用依次转录的方法控制多蛋白结构的组装。A. 柄杆菌鞭毛组装所必需基因的表达谱被标示在鞭毛示意图旁边。左右两旁箭头分别代表鞭毛组装时间和鞭毛基因表达时间，二者是同步的（见正文）。B. 纤毛生成基因的表达谱被标示在纤毛示意图旁边。这些纤毛基因的表达顺序暗示，同鞭毛基因一样，纤毛基因的表达时间可能对纤毛的正确组装起着至关重要的作用。(另见文前彩色插图 8-3)

基因主要在 G1—S 的转型期表达，之后，与染色体开环（decatenation）和分离（segregation）有关基因到表达高潮。总之，这些表达谱都表明柄杆菌起码部分是通过转录调控而在时间上实现对 DNA 复制各个方面的控制。

从这些受细胞周期控制基因的表达谱可以看出，在细胞周期中，柄杆菌细胞一般是在基因被需要的时候，或是比之稍早一些时候表达这些基因。用微阵列分析酵母和哺乳动物细胞的细胞周期中的基因表达谱发现，真核生物也是通过类似的途径来控制细胞周期[14~16]。例如，在酿酒酵母（*Saccharomyces cerevisiae*）中，与 DNA 复制起始有关的基因也是在 G1 期表达量最高，刚好在 S 期开始之前。尽管这些基因与柄细胞的基因没有明显同源性，两种微生物都是在 G1 期开始诱导 DNA 复制起始基因，很显然，这一

机制已经在进化过程中被选择为管理细胞周期的优势机制。

众所周知，基因都是在被需要时或比之稍早时得以表达，依此对受细胞周期控制基因的功能进行分析后发现，另外还有一些细胞过程可能也依赖于细胞周期的进展，其中包括三组编码核糖体、RNA 多聚酶和氧化呼吸中烟酰胺腺嘌呤二核苷酸（nicotinamide adenine dinucleotide，NADH）脱氢酶复合体亚基的基因。这三组基因大约都是在同一时间（紧随 G1—S 转型期之后）被诱导，表明柄细胞和前分裂期细胞比游动细胞有更高的代谢需求，还表明向有复制能力状态下的转变会伴随向代谢活跃状态的总体转变。

通过序列分析无法预测其功能的那些基因，野生菌的基因表达谱可能有助于给它们指定（assign）基因功能。有约 260 个依赖细胞周期表达的基因没有被序列分析预测出它们的功能，分析这些基因的表达时间并分析与这些基因一起受共同机制调控其他基因的功能，有助于准确预测这些未知基因的功能，起码可以建立一些通过实验来验证的假说（testable hypotheses）。例如，有 10 个在游动细胞和柄细胞早期表达的基因被序列分析预测为编码细胞壁合成酶。已经知道，从游动细胞向柄细胞转变时，柄的生成需要在细胞一端定点合成细胞囊。因此，从逻辑上可以推断，在这个从 G1 期到早 S 期转变过程中被诱导与细胞壁代谢有关的一套基因很可能参与了柄的合成。

为什么基因要在不同时期表达

我们的研究经总结发现，有 550 多个依赖细胞周期表达的转录产物，因此，提出了从前很大程度上没有解答的一个重大问题：为什么这些基因要在细胞周期的不同阶段表达？正如前面讨论鞭毛和纤毛的生物合成以及细胞分裂机制时提到过的，转录上的时间调控可能与大亚细胞结构的协调装配有关。另外，有些基因如果在错误时间表达，其产物会对细胞造成伤害。例如，柄杆菌 *ccrM* 基因编码一种必需的 DNA 甲基转移酶，在错误时间表达该基因，或该基因持续表达时，会产生极不健康的细胞，说明控制其转录时间对把该基因的活性精确地限制在细胞周期的某个阶段起着至关重要的作用[17]。仅仅在正需要该基因产物或稍早时才转录和翻译（推测是这样）该基因，可使细胞节约能量和资源，这可能是另外一个驱动转录时间调控的因素。这种生物系统中“赶巧（just-in-time)”的设计原则会提高细胞的运行效率，促进细胞健康生长，从而增强它们的长期适应能力。

柄杆菌利用细胞周期来调节转录的另外一个原因，可能是借此平衡（offset）细胞分裂时，某些基因产物的不对称分布——柄杆菌不对称分裂时的一种常见现象[18~20]。组氨酸激酶 CckA，从蛋白质水平上来看，它在整个细胞周期中都维持一个大约恒定的水平[20]。在荧光显微镜下观察 GFP 标记的 CckA 发现，CckA 在细胞分裂后主要存在于游动细胞中。有趣的是，*cckA* 的转录是随细胞周期变化而变化，它的 mRNA 量在柄细胞中达到最高峰[9]。在柄细胞中增强 *cckA* 的转录可能会平衡 CckA 蛋白在游动细胞中大量分配所带来的影响，从而在蛋白质水平上维持 CckA 的恒定，正如在蛋白质印迹(Western blot）实验中所观察到的那样。组氨酸激酶 PleC 也存在于所有细胞类型中，但在细胞即将分裂时主要存在于游动细胞的那一端，像 *cckA* 那样，*pleC* 的 mRNA 也在柄细胞中急剧增加。最后，基因在特定细胞周期阶段的转录可能与其基因产物的局部

定位（localization）相耦联。出芽酵母在某些情况下，芽颈蛋白（bud-neck protein）只有在某一特殊的细胞类型中才能定位，因为其定位过程（需要将该蛋白直接分泌到芽颈中）要求从头开始（*de novo*）转录并产生该蛋白[21]。因此，酵母蛋白质的正确定位要求基因在某些特定细胞类型中或在细胞周期的某些阶段进行转录，类似情况也存在于柄杆菌中，使某些基因的表达受细胞周期控制。

剖析调控网络

在发现依赖细胞周期表达的基因中，有 34 个双组分信号传导基因（two-component signal transduction gene）和 5 个结合 RNA 聚合酶的 σ 因子。据推测，既然这些基因依赖于细胞周期而表达，受它们调控的基因也应该是依赖细胞周期而表达的。我们尤其对双组分信号传导基因感兴趣，它们中有些基因已经被证实在控制细胞周期的进程中起关键作用。双组分信号传导基因由回应调节子（response regulator）和组氨酸激酶组成，它们是细菌中胞内信号传导的主要形式，也存在于植物和其他一些真核生物中[22]，双组分信号传导蛋白最广为人知的功能是对环境的改变做出适应性反应。但是，柄杆菌的一个回应调节子 *ctrA* 却被证实在控制细胞周期的进程中起关键作用[23]。

最初发现 *ctrA* 基因是由于一个对温度敏感、失去功能的等位基因 *ctrA401*^{ts}[23]，在允许温度（permissive temperature）30℃下，含有 *ctrA401*^{ts}的突变株能正常生长，并能保持完整的细胞形态，但是，如果把它转移到限制性温度（restrictive temperature）37℃下，该突变株就会失去运动能力，不能生成柄，而细胞形态呈现长丝状。此外，*ctrA401*^{ts}突变株在限制温度下会失去存活力，表明 *ctrA* 基因是必需的。CtrA 控制许多与细胞周期变化有关基因的表达，如 DNA 甲基化基因、细胞分裂基因和鞭毛合成基因[23]。此外，CtrA 还能结合复制起点区（origin of replication）的至少 5 个位点，直接抑制 DNA 复制的起始[24]。

知道了 CtrA 在控制细胞周期所起的关键作用，就不难理解细胞在多个水平上对 CtrA 的调控（图 4）。在从游动细胞向柄细胞转变的过程中，CtrA 蛋白被迅速降解（这一过程依赖于 ClpXP 蛋白酶[25]），从而使 DNA 得以复制，随后，柄细胞便开始抑制 CtrA 的降解，慢慢积累新合成的 CtrA。P1 启动子开始大量转录，首先产生一部分 CtrA，随着量的增加，CtrA 开始抑制 P1 启动子的表达，却大大激活了 P2 启动子的表达[26]。这种正反馈（positive feedback）使 CtrA 水平在分裂前母细胞中快速增长，CtrA 只有在被磷酸化以后才具有调控因子的功能，与 CtrA 蛋白表达趋势一致，它的磷酸化趋势也随细胞周期的变化而变化[18,27]。即使设法使 CtrA 在从游动细胞向柄细胞转变时不被降解，它的磷酸化趋势，即它的活性仍然受细胞周期的调控[18]。

CtrA 是细胞的主调节子（master regulator），最近用整体性分析对受其控制的基因和细胞周期变化进行了研究。首先，用 DNA 微阵列找出了所有依赖 CtrA 进行正常表达的基因[9,28]，在允许温度 30℃下培养 *ctrA401*^{ts}突变株，收集其 RNA 样品。然后再把培养温度切换到限制温度 37℃，继续培养 4 小时后，即在细胞将死亡前，再次收集其 RNA 样品，将两次 RNA 样品在微阵列上进行比较，除去那些仅仅与温度变化有关的基因之后，得到约 200 个在野生细胞中依赖细胞周期表达并受 CtrA 调控的基因。

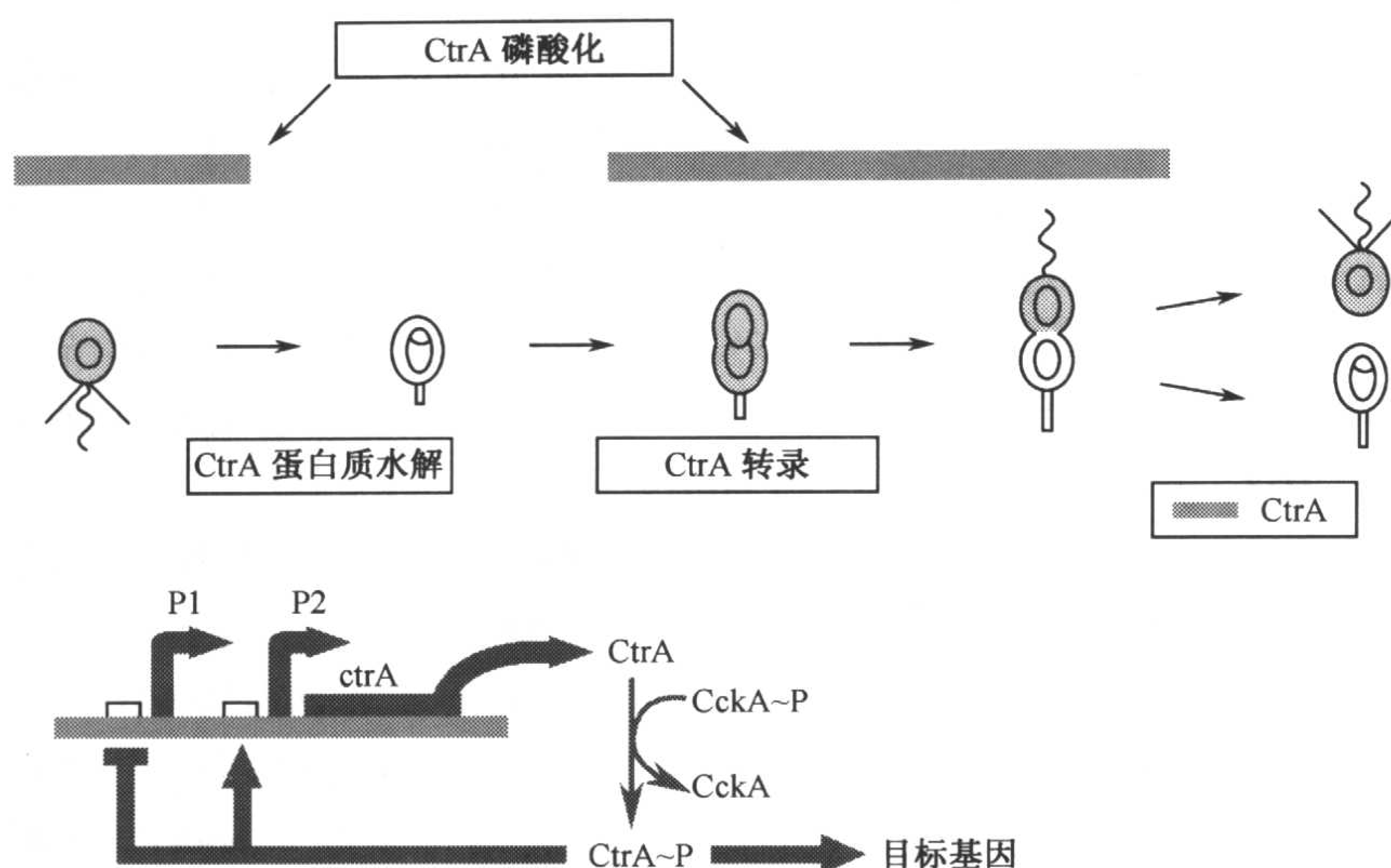


图 4　细胞在不同水平上对 CtrA 在细胞周期中的活性进行调控。CtrA 活性的量在细胞周期中至少被三种机制调控，正文中有详细叙述。图下方示意图表示 *ctrA* 基因转录由 P1 和 P2 两个启动子驱动。柄细胞中，P1 被激活，开始产生 CtrA，当 CtrA 的量积累到一定程度，就开始对 P1 实行负反馈，而对 P2 实行正反馈，使 CtrA 的量快速增加。在柄细胞和前分裂期细胞的柄细胞那一端，CtrA 被水解而不能结合复制原点，从而使 DNA 可以在柄细胞中开始复制。有 CtrA 蛋白的细胞类型被涂成灰色。CtrA 蛋白的一个突变体 CtrAΔ3Ω 可以在细胞中持续表达，而且不被水解，但是依赖于细胞周期的磷酸化过程（在图上方以深灰色条带表示）也使 CtrA 活性随细胞周期的变化而变化。

但是，仅靠表达变化实验并不能指出哪些基因直接受 CtrA 调控，哪些间接受 CtrA 调控。可以试一试一种办法，那就是看预测这些基因[9]的上游调控序列中，有没有与已知 CtrA 结合位点一致的序列。但是不管用哪种算法（algorithm），都不是令人满意，即使有的位点与 CtrA 结合位点完全一致，它们在体内环境中也未必结合 CtrA；相反，CtrA 有可能在体内环境中结合那些与已知结合序列不一致的位点。另外还有一条途径，即“位点分析（location analysis)”[29,30]，可以在全基因组范围内，通过实验把类似 CtrA 这样的转录调控因子的体内结合位点找出来[28]。

在 CtrA 的位点分析实验中，首先，把甲醛加到柄杆菌对数中期（mid-log phase）的细胞液中，这样就使 CtrA（以及其他所有 DNA 结合蛋白）和它在体内所结合的 DNA 链交联（crosslink）在一起。然后，把这些交联好的染色体 DNA 打碎，再用抗 CtrA 的抗体把结合了 CtrA 的 DNA 片段通过免疫沉淀（immunoprecipitation）的方法加以富集。在解除交联后，就可以把富集的 DNA 样品和未富集的 DNA 样品一起用来对 DNA 微阵列进行竞争性杂交；这里所用的微阵列含有代表柄杆菌每一个基因间隔区（intergenic region）的序列斑（spot）。在所富集的 116 个结合 CtrA 的基因间隔区中，可找出 55 个依赖细胞周期和 CtrA 进行正常表达的邻近基因或操纵子，这些基因就组成了 CtrA 细胞周期调控元（regulon）。由于一个操纵子由多个基因组成，所以这个调控元总共有 95 个直接受 CtrA 控制的基因，除了一个基因被遗漏外，这里包括了其他所有在此

前通过其他实验已经证实、直接受 CtrA 控制的基因。

这个 CtrA 细胞周期调控元所包括的基因大致可分为两种功能：细胞顶端的形态发生（polar morphogenesis）和必需的细胞周期过程（essential cell cycle process）。许多与鞭毛和纤毛形成有关的基因及与趋化作用（chemotaxis）有关的基因都直接受 CtrA 调控，当 CtrA 的量在分裂前母细胞中积累到较高水平后，这些受控制基因的表达也都达到了高潮。除了激活这些与细胞末端形态发生有关的基因外，CtrA 还调节许多必需基因的转录，其中包括 DNA 甲基化酶基因 *ccrM* 和 5 个在细胞壁合成和细胞分裂中不可缺少的基因：*ftsZ*，*ftsQ*，*ftsA*，*ftsW* 和 *murG*。

此外，CtrA 还激活 *clpP* 基因的表达，*clpP* 基因编码必要蛋白酶 ClpXP 的一个组成部分，蛋白酶 ClpXP 能够在特定细胞发育阶段降解 CtrA[25]。回应调节子 DivK 在游动细胞向柄细胞转变的过程中也参与 CtrA 的降解；它的表达也需要 CtrA[25a]。以上这些数据表明，在分裂前母细胞的晚期，当 CtrA 的量积累到较高水平时，有可能通过激活 *clpP* 和 *divK* 基因而启动了自身降解过程。但是，这个模型（model）并不能解释为什么在分裂前母细胞的晚期，CtrA 只在柄细胞的那一半被降解，一定还有其他机制（可能与转录无关）在空间上控制 CtrA 的水解。抛开空间因素不谈，这个细菌细胞周期的调节子要对自己的衰亡负责，这与真核生物一些关键的细胞周期调节子类似，如酿酒酵母的周期蛋白（cyclin）Clb2[16,31,32]。

CtrA 细胞周期调控元，包括至少 14 个调控基因，这些基因有可能控制别的细胞周期变化并最终把其他途径连接到细胞周期调控总网络的由 CtrA 控制的子网络（CtrA-controlled subnetwork）中（图 5）。为了把推动柄杆菌细胞周期正常进行的调控网络搞清楚，还需要对其余的调节子的表达变化和整体结合位点进行分析。在酿酒酵母中，已经对 12 个细胞周期调节因子进行了类似的系统研究[33]。对这 12 个因子进行位点分析后发现，它们组成了一系列相互连接、包括多种转录激活蛋白的环路调控途径，从而有助于酵母细胞周期的正常轮回。在柄杆菌的调控网络中，也可能有类似的环路调控途径。

除了那 55 个直接受 CtrA 控制并依赖细胞周期表达的基因或操纵子以外，还有差不多相同数目的基因依赖 CtrA 进行正常表达，但是，在整体位点分析实验中却不结合 CtrA，其中可能有很多基因间接受 CtrA 调控（图 5）。例如，失去功能的 CtrA 会对编码核糖体蛋白的基因和编码 NADH 脱氢酶复合体的基因产生很大的影响，但是，这些基因上游的基因间隔区却不能在 CtrA 的免疫沉淀实验中得到富集。

总之，在推动柄杆菌细胞周期正常运行的遗传网络中，CtrA 是个主要的节点（hub）（图 5），这个主调节子控制着惊人数量的基因，协调着一连串细胞周期和形态发生过程。尽管对 CtrA 调控作用的认识有了相当大的提高，但还是遗留了很多问题：在覆盖了半个细胞周期的时间里，CtrA 如何在全然不同时间段中激活基因的表达？虽然柄杆菌中有四分之一依赖细胞周期表达的基因直接或间接地受 CtrA 调控，谁控制剩余的那四分之三（约 350 个基因）独立于 CtrA 的基因呢？

要回答这些问题，可能还要涉及其他一些未知的调控分子，这些分子本身可能受细胞周期调控。正如前面提到的，在 550 多个受细胞周期调控的基因中，有 34 个双组分信号传导基因，5 个 σ 因子及 10 个其他转录因子。所有这些基因都有可能控制着其他

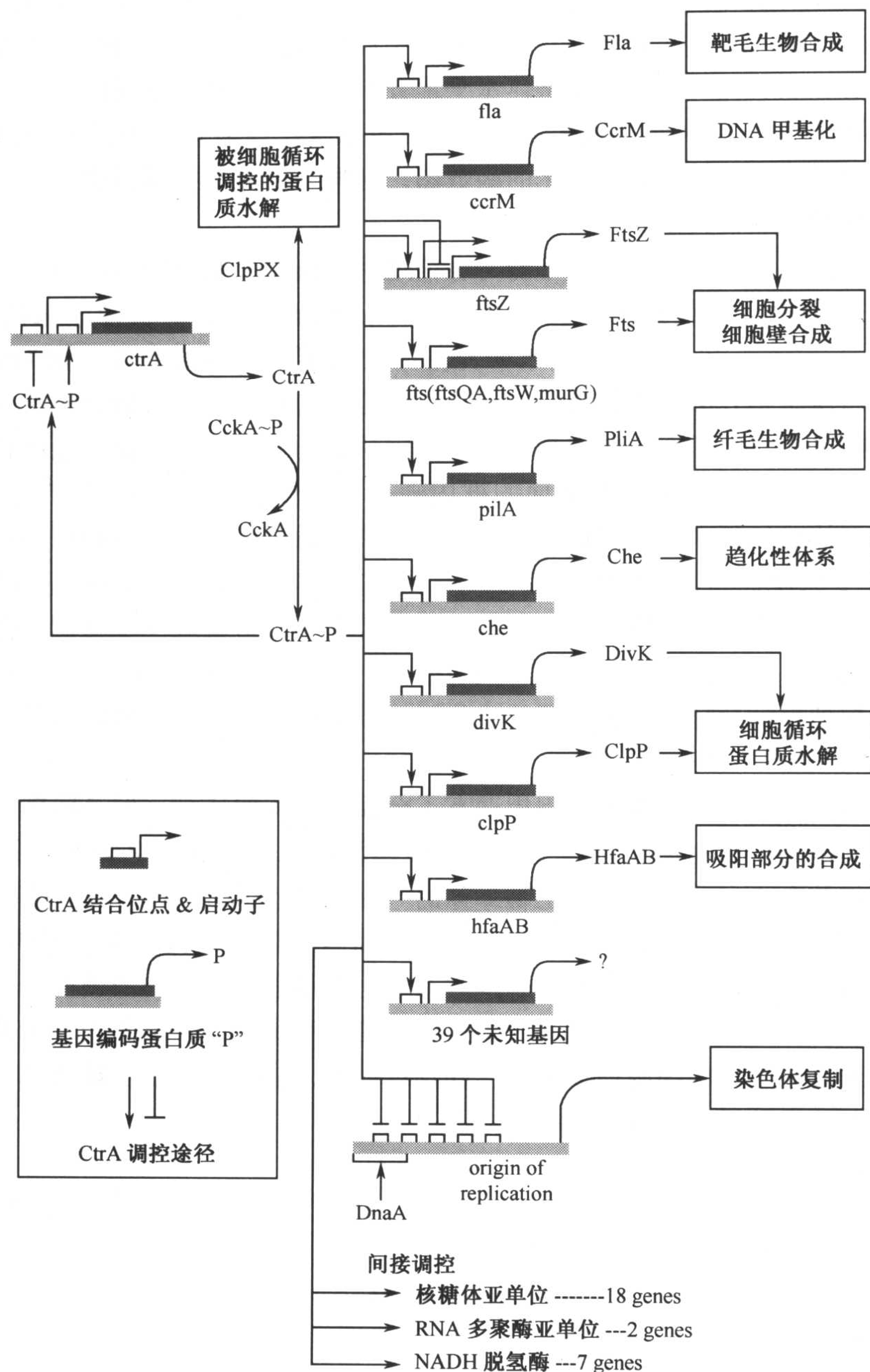


图 5 受 CtrA 控制的基因和细胞周期变化。CtrA 作为转录调控因子直接控制一大批细胞周期变化（在图右边的方框中列出）。正如正文中所描述的那样，DNA 微阵列能很快发现 CtrA 调控元。

一系列依赖细胞周期而表达的基因，对它们的深入研究正在进行中。

除转录变化外，尽管 DNA 微阵列在研究细菌细胞周期活动中，在短期内提供了很多新数据，加深了对它的认识。但是，该技术也存在着明显的缺陷，首先，还有其他水平的调控机制协助控制细胞周期，而这些机制却无法用 DNA 微阵列来研究。其中包括

转录后修饰，如磷酸化、控制蛋白质水解、调控蛋白质之间的相互作用及亚细胞水平上的定位，已经知道所有这些修饰都在柄杆菌细胞周期调控中的某些特定环节起作用。

最近，有一项研究已经开始在整体水平上分析记录细胞周期中蛋白质的合成及其稳定性[34]，以柄杆菌细胞周期为自变量，以蛋白质水平、磷酸化状态以及蛋白质的定位作为变量，可以绘出它们随时间变化的函数图，这样一套完整数据对建立一个综合的柄杆菌细胞周期模型非常关键。

即使收集到这样一套数据，也不能完全描述柄杆菌中细胞周期调控的各个方面，还缺少一个关键部件，那就是用遗传干扰（genetic perturbation）和接下来的表型鉴定（phenotypic characterization）的方法来研究基因的功能。整体研究法，如 DNA 微阵列和大多数蛋白质组学方法，都必须最终和基因组范围内基因功能的研究相结合。

已经在酵母中建立了在整体水平上研究基因功能的技术：大规模、近饱和的插入突变（insertional mutagenesis）[35]和全面的框内缺失（in-frame deletion）[35~37]。酵母框内缺失的突变群体已经用来鉴别酵母在基本培养基中生长时，完成其细胞周期所不可缺少的基因[37]。令人吃惊的是，这些基因种类与从营养丰富培养基中得到的基因种类没有相关性，从细胞周期不同阶段得到的基因种类各不相同，这就近一步强调了把整体功能分析（global analyse of function）和整体表达分析（global expression analyse）相结合的必要性。要想把推动柄杆菌细胞周期的遗传网络系统地描绘出来，最终需要把这些技术结合起来。

基因组学使我们逐渐对细菌细胞周期有了全面、整体的认识，但是问题和挑战也随之而来。尤其是对真核生物和原核生物细胞周期调控机制的相同之处和不同之处感兴趣，例如，细菌会不会采用像真核生物那样的“检查（checkpoint）”和“监视（surveillance）”机制来控制细胞周期变化[38]？各种胞内和胞外信号如何影响细胞周期的进行？细胞如何把形态变化和细胞周期的机制耦联在一起？尽管两界生物的调控基因在序列上的保守性不高，但它们的设计原则和调控机制可能具有保守性。事实上，借助基因组学方法把细菌和真核生物的细胞周期放在一起加以研究，会更有利于发现这些在进化过程中被一再保留的设计原则和遗传结构。

（许朝晖　译）

参考文献

1. Nierman WC, Feldblyum TV, Laub MT, et al. Complete genome sequence of *Caulobacter crescentus*. Proc Natl Acad Sci USA 2001; 98:4136–4141.
2. Fukuda A, Iba H, Okada Y. Stalkless mutants of *Caulobacter crescentus*. J Bacteriol 1977; 131: 280–287.
3. Levin PA, Shim JJ, Grossman AD. Effect of *minCD* on FtsZ ring position and polar septation in *Bacillus subtilis*. J Bacteriol 1998; 180:6048–6051.
4. Marston AL, Thomaides HB, Edwards DH, Sharpe ME, Errington J. Polar localization of the MinD protein of *Bacillus subtilis* and its role in selection of the mid-cell division site. Genes Dev 1998; 12:3419–3430.

5. Marston AL, Errington J. Selection of the midcell division site in *Bacillus subtilis* through MinD-dependent polar localization and activation of MinC. Mol Microbiol 1999; 33:84–96.
6. de Boer PA, Crossley RE, Rothfield LI. A division inhibitor and a topological specificity factor coded for by the minicell locus determine proper placement of the division septum in *E. coli*. Cell 1989; 56:641–649.
7. Newton A. Role of transcription in the temporal control of development in *Caulobacter crescentus*. Proc Natl Acad Sci USA 1972; 69:447–451.
8. Gober JW, England JC. Regulation of flagellum biosynthesis and motility in Caulobacter. In: Brun YV, Shimkets LJ (eds). Prokaryotic Development. Washington, DC: ASM Press, 2000, pp. 319–39.
9. Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L. Global analysis of the genetic network controlling a bacterial cell cycle. Science 2000; 290:2144–2148.
10. Kalir S, McClure J, Pabbaraju K, et al. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science 2001; 292:2080–2083.
11. Skerker JM, Shapiro L. Identification and cell cycle control of a novel pilus system in *Caulobacter crescentus*. Embo J 2000; 19:3223–3234.
12. Sackett MJ, Kelly AJ, Brun YV. Ordered expression of *ftsQA* and *ftsZ* during the *Caulobacter crescentus* cell cycle. Mol Microbiol 1998; 28:421–434.
13. Wortinger M, Sackett MJ, Brun YV. CtrA mediates a DNA replication checkpoint that prevents cell division in *Caulobacter crescentus*. Embo J 2000; 19:4503–4512.
14. Cho RJ, Campbell MJ, Winzeler EA, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998; 2:65–73.
15. Cho RJ, Huang M, Campbell MJ, et al. Transcriptional regulation and function during the human cell cycle. Nat Genet 2001; 27:48–54.
16. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 1998; 9:3273–3297.
17. Zweiger G, Marczynski G, Shapiro L. A *Caulobacter* DNA methyltransferase that functions only in the predivisional cell. J Mol Biol 1994; 235:472–485.
18. Domian IJ, Quon KC, Shapiro L. Cell type-specific phosphorylation and proteolysis of a transcriptional regulator controls the G1-to-S transition in a bacterial cell cycle. Cell 1997; 90: 415–424.
19. Wheeler RT, Shapiro L. Differential localization of two histidine kinases controlling bacterial cell differentiation. Mol Cell 1999; 4:683–694.
20. Jacobs C, Domian IJ, Maddock JR, Shapiro L. Cell cycle-dependent polar localization of an essential bacterial histidine kinase that controls DNA replication and cell division. Cell 1999; 97:111–120.
21. Lord M, Yang M C, Mischke M, Chant J. Cell cycle programs of gene expression control morphogenetic protein localization. J Cell Biol 2000; 151:1501–1512.
22. Loomis WF, Kuspa A, Shaulsky G. Two-component signal transduction systems in eukaryotic microorganisms. Curr Opin Microbiol 1998; 1:643–648.
23. Quon KC, Marczynski GT, Shapiro L. Cell cycle control by an essential bacterial two-component signal transduction protein. Cell 1996; 84:83–93.
24. Quon KC, Yang B, Domian IJ, Shapiro L, Marczynski GT. Negative control of bacterial DNA replication by a cell cycle regulatory protein that binds at the chromosome origin. Proc Natl Acad Sci USA 1998; 95:120–125.
25. Jenal U, Fuchs T. An essential protease involved in bacterial cell-cycle control. Embo J 1998; 17:5658–5669.
25a. Hung DY, Shapiro L. A signal transduction protein cues proteolytic events critical to *Caulobacter* cell cycle progression. Proc Natl Acad Sci USA 2002; 99:13,160–13,165.

26. Domian IJ, Reisenauer A, Shapiro L. Feedback control of a master bacterial cell-cycle regulator. Proc Natl Acad Sci USA 1999; 96:6648–6653.
27. Quon KC. Thesis: Temporal control during the *Caulobacter crescentus* cell cycle [doctoral thesis]. Stanford, CA: Stanford University, 1996.
28. Laub MT, Chen SL, Shapiro L, McAdams HH. Genes directly controlled by CtrA, a master regulator of the *Caulobacter* cell cycle. Proc Natl Acad Sci USA 2002; 99:4632–4637.
29. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 2001; 409:533–538.
30. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. Science 2000; 290:2306–2309.
31. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ. Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol Biol Cell 2000; 11:369–391.
32. Yeong FM, Lim HH, Padmashree CG, Surana U. Exit from mitosis in budding yeast: biphasic inactivation of the Cdc28-Clb2 mitotic kinase and the role of Cdc20. Mol Cell 2000; 5:501–511.
33. Simon I, Barnett J, Hannett N, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell 2001; 106:697–708.
34. Grunenfelder B, Rummel G, Vohradsky J, Roder D, Langen H, Jenal U. Proteomic analysis of the bacterial cell cycle. Proc Natl Acad Sci USA 2001; 98:4681–4686.
35. Ross-Macdonald P, Sheehan A, Roeder GS, Snyder M. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 1997; 94:190–195.
36. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet 1996; 14:450–456.
37. Winzeler EA, Shoemaker DD, Astromoff A, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 1999; 285:901–906.
38. Hartwell LH, Weinert TA. Checkpoints: controls that ensure the order of cell cycle events. Science 1980; 246;629–634.

第四部分：微生物基因组的进化

9

原核生物的进化及分类简史

Lorraine Olendzenski, Olga Zhaxybayeva and J. Peter Gogarten

引言

现在正值原核生物系统学的新纪元，开展研究具备了阐明所有生物，特别是原核细菌和古生菌自然进化关系的手段和知识。这并不一定正确，也许30年后，可能会对这些生物的种系发生关系有不同认识。毫无疑问，日益宠大的基因组数据库将有助于认识原核生物自然关系，并正影响着目前对原核生物进化的理解。

接近原核生物自然系统：早期对其成员及关系的描述

如果种系发生学是为了提出一个系统，即通过该系统能反映物种间的自然关系并重建它们的世代关系，由于历史原因，原核微生物的这一目的一直是个难题。在用遗传学方法研究原核生物系统学之前，最早是建立在从光学显微镜所能获得的有限数据上。自列文·虎克（Leeuwen Hoek）发现微生物世界后，早期的研究者们试图给这些观察到的生物命名和描述，考虑到当时已有的名称和描述，几乎没有人考虑它们的自然关系（种系发生），尽管假设可能有一个细菌自然系统，但仍采用与动物和植物命名相同的结构(即目，科，属等)[1]。

由于显微镜是主要观察工具，因此大部分描述只是基于形态、行为和栖息地，而且当时主要观察水生环境[2,3]。Haeckel[4]建立了一个在原生生物（Protista）下的新亚群，单体（Monere)。Cohn[5~7]研究的独特之处在于，他认识到所观察这些不同形态和大小的生物或许代表不同的种类，而不是同一生物的不同生命时期。Cohn 认为 Ehrenberg 和 Dujardin 的早期研究是新细菌分类的有效基础[2,8,9]。根据细胞形态，他把细菌分为4群，并认为细菌（如当时所知）通过蓝绿藻与植物相关，它们应划分在一起。20世纪50年代前，Orla-Jensen 提议用生理特征，如代谢副产物、糖发酵、生长温度范围来建立一个广泛、基于进化关系的分类系统[3,10]。

1941年，Stanier 和 van Neil[11]为单体（Monera）界提出了一个基于种系发生的分类系统，它是由 *Myxophyta*（蓝细菌）和裂殖菌（*Schizomycetae*，所有其他细菌）组成。裂殖菌下包含3个主要的群［真细菌，包括光合细菌，放线菌和众多的单细胞细菌；黏细菌（*Myxobacteriae*）和螺旋体（*Spirocheatae*)］。他们还提出了其他许多补充的附属群，如包括纤丝菌（*Leptothrix*)、齿丝菌（*Crenothrix*)、无色菌（*Achromatium*)、巴斯德菌（*Pasteuria*）和丝微菌（*Hyphomicrobium*）群。细胞形状和鞭毛的排列被作为重要的分类特征。依照原先的观点，无色的硫氧化菌（*Beggiatoa*）与蓝细菌（*Myxophyta*）分在一起，这是因为他们明显相似的形态特征[11]。

Stanier 和 van Neil[11]反对用生理学作为种系发生特征，因为这会把根据形态学特征划分在一起的自然群分开，也会将那些不应分在一起的生物强分在一起。确定哪些生理特征是先天的还是后天的也成问题。尽管他们长期从事细菌进化关系方面的研究，但是直到 1946，Van Neil 改变了他的想法，放弃了所有向种系发生方面的努力[12]，普遍认为，由于缺乏分类特征，特别是缺乏可以通过光学显微镜观察到的形态特征，使原核生物的种系学研究不可能。

原核生物分类：经验方法

在现代分子种系发生方法出现之前，大部分原核生物分类系统是由主观确定的，仅仅就是为了给某种生物命名，没有试图说明它们的进化关系。Cherter 的早期研究极大的影响了“伯杰氏生物鉴定手册”第一版的成形，他认为，没有系统分类单元的描述和分类系统，就无法鉴定和确认细菌新种[13]。这本手册，被大多数研究者奉为细菌分类学“经典”，成功出版到第九版，后又改名为“伯杰氏系统细菌学手册”，现已是第二版了。

“伯杰氏生物鉴定手册”1923 年由美国生物学家协会（The Society of American Biologist）发起，随后由 1936 年成立独立的伯杰氏手册委员会（Bergey's Manual Trust）指导下出版。在国际微生物协会（International Association of Microbiological Science）指导下，细菌分类学法规随着国际系统细菌学委员会（International Committee on Systematic Bacteriology）的成立而形成[3]，该委员会实施的一些重大改革使现在的原核生物分类与命名得以形成，并公布了一个经批准的细菌名称名单（Approved List of Bacterial Names)[14]，把细菌命名的起始日期定为 1980 年 1 月 1 日，取代了 1753 年 5 月 1 日 [Linneaus 起用植物菌（*plantarum*）这一种名的那天]，规定了命名的有效发表和无效发表的细菌法规（Bacterial Code)[15]，废除了不在批准名单上的名字以备将来使用[3]。

“伯杰氏手册”的漫长历史证实了它作为一本细菌鉴定手册的广泛性和实用性。使用计算机分析大量表型数据的出现（数字分类或 Adansonian 方法[16]），提高了表型鉴定的有效性并巩固了已确定的方法。所有有用数据（或统计学上显著部分）可用来计算菌株间的相似性，且物种能根据表型的相似性加以分类[17,18]。这些经验方法没有考虑到相似性的起源[19]，因此并不具有种系发生学意义，通过水平基因转移而获得的性状容易把物种划分在一起，从而模糊了它们的进化史。

细菌的概念

当“伯杰氏”建立的时候，细菌细胞的特性并没有完全弄明白，细菌（裂殖菌）分类学讨论常在植物学聚会上。Stanier 和 van Neil[20]意义深远的研究把细菌定为一个独特的群。随着电子显微镜的应用，原核生物和真核生物细胞的根本区别更清晰，细菌（按当时的定义）缺少核膜和细胞器，有明显细胞壁。细菌最初是通过排除法定义的，即通过定义没有的特征而不是一系列独特已有的特征[21]。在许多年后发现古生菌时，对细菌的定义（根据已有的特征而不是根据缺乏的特征）就有意义多了。

遗传学和非遗传学方法：分类学和种系发生学的新工具

随着对DNA是细胞遗传物质并控制着表现性状认识的加深，分子生物学和遗传学方法应用到物种的种系发生学关系中。这激起了一些微生物学家用遗传学方法研究细菌进化分类的新兴趣[22]。DNA碱基组成，即鸟膘呤和胞嘧啶（G+C）的摩尔百分比提供了一个粗略的分类依据，从而有助于确定关系相近的物种，G+C百分比不同可以明确地把物种划分为不同的群，尽管今天没有把它作一个主要分类指标，碱基组成仍然包括在对原核生物类群的划分之中。

核酸杂交技术，通过测定热变性温度，也是对物种间总体相关性程度的一种粗略量化表示[22,24]。两种不同细菌的单链DNA温育在一起能形成异源的双链，该异源双链中未形成配对的碱基百分比反映了两种菌的异源程度，未配对的碱基数目可以粗略地用同源双链与异源双链DNA热稳定性的差值（ΔT_m）来表示，如果不配对的碱基超过20%，双链就不会形成[18]。蛋白质的免疫学比较，即定量确定蛋白质的交叉反映也有助于划定近缘物种[19,22]。Arnheim及其同事的研究支持这种遗传物质相关性的推测[25]，并证明免疫学相似性与序列同源性有关。

所有这些技术解决了许多较低水平的种系发生问题（例如，种，科，门），但都不能确定最高分类水平之间的关系或者推断门与门之间的关系。这些进展促使Stanier预料到未来的结果将是一个“有断层的金字塔（fragmented heirarchy）”，一个基于遗传学的分类系统，其中已知同源类群将在更高水平上保持分散，然而仍被归在一起作为原核生物[22]。

原核生物的种系发生：它的分子基础

Zuckerkandl和Pauling[26]认为，生命的历史可以记录在核甘酸和蛋白质的序列之中，这为微生物分类学革命打下了基础。第一个用于种系进化树的分子是细胞色素C和铁氧还原蛋白的氨基酸序列[27,28]，Schwartz和Dayhoff[29]用它们来构建原核生物和真核生物的种系发生关系，通过这些数据强化了这种观点，即真核生物的细胞器、线粒体和质体来源于自由生活的细菌[30]。

用RNA寡核苷酸分子编目法，Fox，Woese及其同事[31~33]开创了种系发生方法的新纪元，并提供了可将所有生物进行比较的尺度。小亚基RNA分子分布广泛，在所有生物中行使同样功能，并与其他许多组分相互作用，它既有高度保守区域，因此不易进行水平转移，又有高度可变区域，因此能在生物之间进行比较[21]。

这项工作刚开始时，还无法测定RNA全序列，因此，通过核糖核酸酶T1消化得到16S RNA的许多小片段，再对它们进行测序，以及分类和类群之间的比较[34]。对RNA编目比较可建成种系发生史，早期对细菌种系发生关系的变化必然导致细菌分类在最高水平上变化。

发现产甲烷菌能形成相互聚类的一群，与先前研究的其他所有细菌都不同且亲缘关系较远[35]。同时，对产甲烷菌和嗜盐菌的研究发现，它们的细胞壁中不含肽聚糖，这

也与其他已知细菌不同[36]，从而导致将原核生物界划分为两种根本不同类型的细胞：真细菌（*Eubacteria*）和古生菌（*Archaebacteria*），即现在众所周知的细菌（*Bacteria*）和古生菌（*Archaea*）[33,37,38]。

后来，测序全长 rDNA 分子［即编码核糖体 RNA（rRNA）的 DNA］成为常规技术，从而允许对生物大量的序列信息进行比较，结合其他一系列技术推断出种系发生史[39]。扩增 rRNA 编码基因的手段（直接克隆环境中的 DNA 或用聚合酶链反应），使得不仅可以分析培养物种的 rRNA，还可以分析环境样品中的 rRNA[40~42]。随着环境中分离越来越多的 rRNA 序列不能与任何已培养生物相符，人们意识到原核生物的多样性被大大低估了[43]，被培养和正式命名的原核生物不足 5000 种，然而，这仅仅占能培养的原核生物 1%～10%[44]。

通常对培养物和环境样品中 rRNA 序列的测定已导致 rRNA 数据库急剧膨胀，而且这种趋势还在继续[45]。一个全面的，包括三个域的种系发生观已被接受：即细菌、古生菌和真核生物[38]。根据这一进展，"伯杰氏系统细菌学手册"采用了一个基于 16S 核糖体（rRNA）系列来确定门的命名方法。第二版列出了古生菌 2 个门和细菌 23 个门[46]（表 1）。用 16S rRNA 方法确定的门与原来严格限定的分类方法确定门的重叠部分将在文献［46］的图 2 中介绍。

表 1 常用细菌门的分类系统

伯杰氏手册[23]	不同门的代表[43, 99]	国家生物信息中心分类数据库，2002 年 1 月[100]
细菌		
BⅠ产液菌(*Aquificae*)	产液菌(*Aquificales*)	产液菌(*Aquificae*)
BⅡ栖热袍(*Thermotogae*)	栖热袍菌（*Thermotogales*）	栖热袍菌(*Thermotogae*)
BⅢ栖热脱硫细菌（*Thermodesulfobacteria*）	栖热脱硫细菌（*Thermodesulfobacteria*）	栖热脱硫细菌（*Thermodesulfobacteria*）
BⅣ高温异常球菌(*Deinococcus-Thermus*)	栖高温/异常球菌（*Thermus/Deinococcus*）	栖高温/异常球菌（*Thermus/Deinococcus group*）
BⅤ产金菌(*Chrysiogenetes*)	—	产金菌(*Chrysiogenetes*)
BⅥ绿屈挠菌(*Chloroflexi*)	绿色非硫细菌(*Green nonsulfur bacteria*)	绿屈挠菌(*Chloroflexi*)
BⅦ栖热微菌(*Thermomicrobia*)	—	栖热微菌(*Thermomicrobia*)
BⅧ硝化螺菌（*Nitrospirae*）	硝化螺菌(*Nitrospira*)	硝化螺菌(*Nitrospirae*)
BⅨ铁还原杆菌(*Deferribacteres*)	联合菌(*Synergistes*) 柔柄菌(*Flexistipes*)	铁还原杆菌(*Deferribacteres*)
BⅩ蓝细菌(*Cyanobacteria*)	蓝细菌(*Cyanobacteria*)	蓝细菌(*Cyanobacteria*)
BⅪ绿菌(*Chlorobi*)	绿菌(*Chlorobiaceae*)	拟杆菌/绿菌群（*Bacteroidetes/Chlorobi* group）
BXX拟杆菌(*Bacteroidetes*)	拟杆菌/噬纤维菌（*Bacteroidetes/Cytophaga*）	—

续表

伯杰氏手册[23]	不同门的代表[43, 99]	国家生物信息中心分类数据库，2002 年 1 月[100]
BⅫ多形杆菌（*Proteobacteria*）	多形杆菌（*Proteobacteria*）	多形杆菌/紫色细菌及相关菌（*Proteobacteria*/Purple bacteria and relatives）
BⅩⅢ厚壁菌（*Firmicutes*）	低 G+C 革兰氏阳性菌	厚壁菌（*Firmicutes*）
BⅩⅣ放线菌（*Actinobacteria*）	放线菌（*Actinomycetales*）	—
BⅩⅤ浮霉状菌（*Planctomycetes*）	浮霉状菌（*Planctomycetales*）	浮霉状菌（*Planctomycetales*）
BⅩⅥ衣原体（*Chlamydia*）	衣原体（*Chlamydia*）	衣原体/疣微菌（*Chlamydiales*/*Verrucomicrobia*）
BⅩⅩⅡ疣微菌（*Verrucomicrobia*）	疣微菌（*Verrucomicrobium*）	—
BⅩⅦ螺旋体（*Spirochaetes*）	螺旋体/钩端螺旋体（*Spirochaetes*/*Leptospira*）	螺旋体（*Spirochaetales*）
BⅩⅧ丝状杆菌（*Fibrobacteres*）	丝状杆菌（*Fibrobacter*）	丝状杆菌/嗜酸菌群（*Fibrobacter*/*Acidobacteria* Group）
BⅪⅩ 嗜酸菌（*Acidobacteria*）	嗜酸菌（*Acidobacterium*）	—
BⅩⅪ梭杆菌（*Fusobacteria*）	梭杆菌（*Fusobacteria*）	梭杆菌（*Fusobacteria*）
BⅩⅩⅡ网球菌（*Dictyoglomus*）	网球菌（*Dictyoglomus*）	网球菌群（*Dictyoglomus* group）
—	—	脱卤拟球菌群（*Dehalococcoides* group）
	来自温泉的环境门类：OP1，OP3-OP10，EM19[a]	
古生菌		
AⅠ嗜泉古生菌（*Crenarchaeota*）	嗜泉古生菌（*Crenarchaeota*）	嗜泉古生菌（*Crenarchaeota*）
AⅡ广域古生菌（*Euryarchaeota*）	广域古生菌（*Euryarchaeota*）	广域古生菌（*Euryarchaeota*）
—	初生古生菌界（*Korarchaeota*）	初生古生菌界（*Korarchaeota*）
—	—	嗜钠古生菌（*Nanoarchaeota*）

[a]每一门类代表一个单独的门。

现在认为，16s rRNA 差异大于 3%，也就是基因组序列相似性小于 70%，即熔解温度（ΔT_m）大于或等于 5℃（最近的综述见文献［44］）即被人为不同的种。然而，在某些情况下，这种方法引起问题。因为 16s rRNA 的保守特性，某些不同的种可能有相同的或相近的 rRNA 序列[47]，而在其他情况下，一个种的菌株也可能有 rRNA 基因上的差异[48,49]。而且，单个的生物含有多个 rRNA 操纵子；少数情况下，它们有着大于 5%的差异性[50~53]。

探寻生命树之根

第一个基于 16S rRNA 的整体树是没有根的。当研究所有生命时，没有一个群能延伸为种系发生树之根。Schwarz 和 Dayhoff[29]曾建议用横向同源基因序列或编码同一蛋白质的复制基因序列解决这个问题，一个横向种系发生能作为另一个外群，将这种方法用到质子泵腺苷三磷酸酶[54]、延伸因子 EF-Tu 和 EF-G[55~57]、信号识别因子[58]以及氨酰转移 RNA（tRNA）合成酶[59]上构建的种系发生树的根，一边是细菌，另一边是古生菌和真核生物。古生菌有一系列类似真菌生物的特征，主要是基因组结构、翻译和转录[60~62]，然而，它的某些基因非常类似于原核生物的特征，从而意味着这枝谱系是杂合型的[63,64]。

分子核心/树的动摇

随着获得分子数据的增加，不同基因树被构建，并与 rRNA 树进行了比较，有些与其一致，有些却不。对这种用不同标记而构建出不一样的种系发生树，有一种解释是因为水平基因转移（horizontal gene transfer，HGT）[65]。遗传信息的交换是微生物进化的一个主要因素[66,67]，操纵子的形成归因于频繁的基因转移[68]，而基因的获得，通常形成基因组的弧岛，这是能使生物占据一个新生态位的重要因素[69,70]。有人建议用一个种内水平基因交换的高频率定义微生物的种，就像定义高等生物的种那样[68,71]，群体研究认为这种推理至少在某些种内是合理的——即那些种内的种系发生树在拓扑结构上不同，种间的种系发生树却比较一致[71,72]。

然而，基因转移不仅限于种间，它甚至能出现在域间。除了偶尔选择的基因外[68]，HGT 能影响管家基因[73]和遗传信息处理功能的基因[67,74~76]，通过分析全序列测定的微生物基因组发现，HGT 的广泛性随处可见[63,77~79]。这一复杂学说认为，信息基因（informational gene）因其基因产物的高度合作性而比操作基因（operational gene）更难以被转移[80]，然而，从来没有一组基因从不进行水平转移[81, 82]。

rRNA 和基因组的镶嵌特性

由于 rRNA 在翻译中的中心作用和核糖体的复杂性，传统观念认为一种生物能利用另一种生物 rRNA 的可能性很低。然而，来源于不同生物的核糖体组分能重新组成有功能的核糖体[83~86]。在实验条件下，一种生物的核糖体操纵子能被另一种生物取代[87]，从同一区域反向转录的 rRNA 操纵子能在同一基因组内共存[50,51]，表现出广泛重组性镶嵌式操纵子具有的功能[51,52,88]。难道这种根据其他基因[89~93]与 rRNA 得到的种系发生之间的一致性，是因为基因组与 rRNA 均具有的镶嵌特性吗[94]？

长期以来，即使是低频率水平基因转移也会导致镶嵌式基因组，而不同镶嵌部分反映了不同的历史，一个分子或基因树与生物进化之间的潜在差异已取得共识，后者常被假设为网状[65,73]。

树、网及世代线

尽管有 HGT 蔓延，在一个非常短时间的间隔里，一种生物的世代线仍能确立，因为大多数基因进行一致的纵向遗传（即在细胞分裂时从母代传到子代细胞），似乎分叉物种之间的转移通常只涉及少数基因的转移[64,80,95]。只要考虑的时间为短间隔，上述生物世代的定义在大多数情况适用，然而，可靠基因组信息只能对现存生物，并不是过去的无限代生物。

从不同分子标记得到一致的种系发生关系，说明该物种的这些基因主要是纵向遗传；然而，水平基因转移本身具有产生一致分子种系发生的潜能（见图 1）；而这些一致的分子种系发生关系并不一定反映生物的种系发生，基因组的镶嵌式特性[81,82]使重建生物世代关系变得困难，在某些情况下是不可能的。

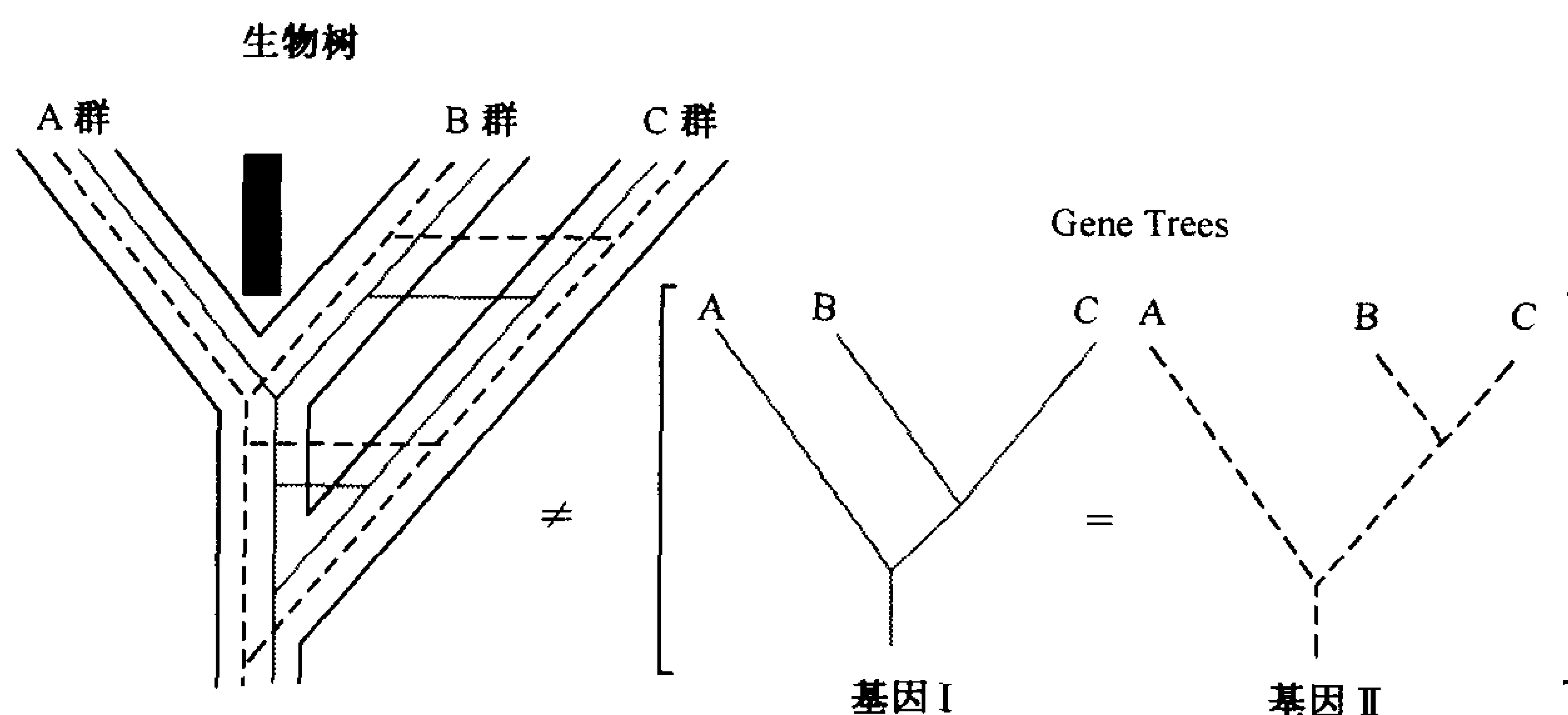


图 1　水平基因转移（HGT）如何影响生物进化的重建。灰线和虚线代表 2 个基因树，包围它们的实线代表生物进化（见正文讨论），A 群是一个更近代的分枝分类元，当它从 B 群分开后，它成为独特的一枝或形成一系列独特的生理特征，使它与群 B 和群 C 相同基因的频率大大减少（竖直的黑条）。另一方面，群 B 与群 C 继续交换基因，结果，独立基因树（Ⅰ和Ⅱ）的分子种系发生关系与物种种系发生关系不同，根据这种情况，基于分子数据的分类更多反映 HGT 的频率而不是纵向基因传递，突变的积累和随后世代分离，尤其是，在基因树里的较长分支世代（A）是因其参与水平基因转移频率较低而不是因其进化得更早。

总结

达尔文关于物种起源的研究[96]以及新达尔文合成论（neo-Darwinian synthesis）[97]认为，在一段长时间内，生物体的进化总能描述为严格的二分枝树状结构，反映生命树的分类学是现代系统学家的目的[98]，刚开始，微生物分类不得不依靠很少的特性，以致不可能建立一个自然分类系统。Woese 和 Fox 引进小亚基 rRNA 作为微生物分类的工具[33]，使许多微生物学家确信关于动物和植物的分类概念能被延伸到原核生物的王国，“直到出现分子测序，细菌进化才可能成为实验方法的对象”[21]，更多的序列数据把物种更准确的放到生命进化树中成为希望。分子数据，特别是 16S rRNA 和全基因组数

据，成为微生物分类有价值的工具。根据分子序列数据确立群的概念，通常反映了生化、生理学和结构的特性。然而，认识到 HGT 是微生物进化的一个重要力量，以及发现微生物基因组的镶嵌特性，引起了对微生物进化的重新认识[94]，分类学的群是反映共同祖先还是倾向于 HGT 仍在争论之中。

致谢

本章的研究由 NASA 空间生物学项目和 NASA 在 Tempe 的亚利桑那州立大学 NASA 航空研究所资助。感谢 Ed Lendbetter，W. Ford Doolittle 和 Otto Kandler 提供有用的文献。

（刘　斌，刘　超　译）

参考文献

1. Müller OF. Animacula Infusoria Fluviatilia et Marina, quae Detexit, Systematice Descripsit et ad Vivum Delineari Curavit, Hauniae: Typis N Mölleri, 1786.
2. Ehrenberg CG. Die Infusionsthierchen als vollkommene Organismen. Leipzig: Leopold Voss, 1838.
3. Murray RGE, Holt JG. The history of Bergey's Manual. In: Garrity GM (ed). Bergey's Manual of Systematic Bacteriology. Volume One. The Archaea and Deeply Branching and Phototrophic Bacteria. New York: Springer, 2001, pp. 1–13.
4. Haeckel E. Generelle Morphologie der Organismen. Berlin: Georg Reimer, 1866.
5. Cohn F. Untersuchungen über Bakterien. Beitr Biol Pflanz 1872; 1:127–224.
6. Cohn F. Untersuchungen über Bakterien II. Beitr Biol Pflanz 1875; 1:141–207.
7. Brock TD. Milestones in Microbiology, 1556 to 1940. Washington, DC: ASM Press, 1998.
8. Dujardin F. Histoire Naturelle des Zoophytes. Paris: Roret, 1841.
9. Drews G. The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. FEMS Microbiol Rev 2000; 24:225–249.
10. Orla-Jensen S. Die Hauptlinien des natürlichen Bakterien-systems. Zentbl Bakteriol Parasitenkd Infektkrankh Hyg Abt II 1909; 22:97–98 and 305–346.
11. Stanier RY, van Neil CB. The main outlines of bacterial classification. J Bacteriol 1941; 42: 437–466.
12. van Neil CB. The classification and natural relationships of bacteria. Cold Spring Harb Symp Quant Biol 1946; 11:285–301.
13. Chester FD. A Manual of Determinative Bacteriology. New York: Macmillan, 1901.
14. Skerman VBD, McGowan V, Sneath PHA. Approved list of bacterial names. Int J Syst Bacteriol 1980; 30:225–420.
15. Lapage SP, Sneath PHA, Lessel J, Skerman VBD, Seeliger HPR, Clark WA. International Code for Nomenclature of Bacteria, 1976 revision. Washington, DC: American Society for Microbiology, 1975.
16. Sneath PHA, Sokal RR. Numerical taxonomy. Nature 1962; 193:855–860.
17. Sokal RR. Typology and empiricism in taxonomy. J Theor Biol 1962; 3:230–267.
18. Brenner DJ, Staley JT, Krieg NR. Classification of procaryotic organisms and the concept of bacterial speciation. In: Garrity GM (ed). Bergey's Manual of Systematic Bacteriology. Volume One. The Archaea and Deeply Branching and Phototrophic Bacteria. New York: Springer, 2001, pp. 27–31.

19. Marmur J, Falcow S, Mandel M. New approaches to bacterial taxonomy. Ann Rev Microbiol 1963; 17:329–372.
20. Stanier RY, van Neil CB. The concept of a bacterium. Arch Mikrobiol 1962; 42:17–35.
21. Woese CR. Bacterial evolution. Microbiol Rev 1987; 51:221–271.
22. Stanier RY. Toward an evolutionary taxonomy of the bacteria. In: Pérez-Miravete A, Peláez D (eds). Recent Advances in Microbiology, 10th International Congress for Microbiology, Mexico: Associacion Mexicana De Microbiologia; 1971, pp. 595–604.
23. Garrity G, (ed). Bergey's Manual of Systematic Bacteriology. New York: Springer, 2001.
24. Marmur J, Doty P. Thermal renaturation of DNA. J Mol Biol 1961; 3:584–594.
25. Arnheim N, Prager EM, Wilson AC. Immunological prediction of sequence differences among proteins. Chemical comparison of chicken, quail, and phesant lysozymes. J Biol Chem 1969; 244:2085–2094.
26. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds). Evolving Genes and Proteins. New York: Academic Press, 1965, pp. 97–166.
27. Margoliash E. Primary structure and evolution of cytochrome c. Proc Natl Acad Sci USA 1963; 50:672–679.
28. Fitch WM, Margoliash E. Construction of phylogenetic trees. Science 1967; 155:279–284.
29. Schwartz RM, Dayhoff MO. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 1978; 199:395–403.
30. Sagan L. On the origin of mitosing cells. J Theor Biol 1967; 14:255–274.
31. Zablen L, Woese CR. Procaryote phylogeny IV: concerning the phylogenetic status of a photosynthetic bacterium. J Mol Evol 1975; 5:25–34.
32. Woese CR, Fox GE, Zablen L, et al. Conservation of primary structure in 16S ribosomal RNA. Nature 1975; 254:83–86.
33. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 1977; 74:5088–5090.
34. Fox GE, Peckman KJ, Woese CR. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. Int J Syst Bacteriol 1977; 27:44–57.
35. Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci USA 1977; 74:4537–4541.
36. Kandler O, Hippe H. Lack of peptidoglycan in the cell walls of *Methanosarcina barkeri*. Arch Microbiol 1977; 113:57–60.
37. Woese CR, Fox GE. The concept of cellular evolution. J Mol Evol 1977; 10:1–6.
38. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990; 87:4576–4579.
39. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (ed). Molecular Systematics. Sunderland, MA: Sinauer, 1996, pp. 407–514.
40. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. Nature 1990; 345:60–63.
41. Schmidt TM, DeLong EF, Pace NR. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 1991; 173:4371–4378.
42. Hugenholtz P, Pace NR. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol 1996; 14:190–197.
43. Pace NR. A molecular view of microbial diversity and the biosphere. Science 1997; 276:734–740.
44. Rossello-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev 2001; 25:39–67.
45. Maidak BL, Cole JR, Lilburn TG, et al. The RDP-II (Ribosomal Database Project). Nucleic Acids Res 2001; 29:173–174.
46. Garrity GM, Holt JG. The road map to the manual. In: Garrity GM (ed). Bergey's Manual of Systematic Bacteriology 2nd ed. New York: Springer, 2001, pp. 119–116.

47. Probst A, Hertel C, Richter L, Wassill L, Ludwig W, Hammes WP. *Staphylococcus condimenti* sp nov, from soy sauce mash, and *Staphylococcus carnosis* (Schleifer and Fisher 1982) subsp utilis subsp nov. Int J Syst Bacteriol 1998; 48:651–658.
48. Fox GE, Wisotzkey JD, Jurtshuk P Jr. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 1992; 42:166–170.
49. Martinez-Murcia AJ, Benlloch S, Collins MD. Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA–DNA hybridizations. Int J Syst Bacteriol 1992; 42: 412–421.
50. Yap WH, Zhang Z, Wang Y. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol 1999; 181:5201–5209.
51. Mylvaganam S, Dennis PP. Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics 1992; 130:399–410.
52. Dennis PP, Ziesche S, Mylvaganam S. Transcription analysis of two disparate rRNA operons in the halophilic archaeon *Haloarcula marismortui*. J Bacteriol 1998; 180:4804–4813.
53. Amann G, Stetter KO, Llobet-Brossa E, Amann R, Anton J. Direct proof for the presence and expression of two 5% different 16S rRNA genes in individual cells of *Haloarcula marismortui*. Extremophiles 2000; 4:373–376.
54. Gogarten JP, Kibak H, Dittrich P, et al. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 1989; 86:6661–6665.
55. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of Archaebacteria, Eubacteria, and Eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 1989; 86:9355–9359.
56. Cammarano P, Palm P, Creti R, Ceccarelli E, Sanangelantoni AM, Tiboni O. Early evolutionary relationships among known life forms inferred from elongation factor EF-2/EF-G sequences: phylogenetic coherence and structure of the archaeal domain. J Mol Evol 1992; 34:396–405.
57. Baldauf SL, Palmer JD, Doolittle WF. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA 1996; 93:7749–7754.
58. Gribaldo S, Cammarano P. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J Mol Evol 1998; 47:508–516.
59. Brown JR, Doolittle WF. Root of the universal tree of life based on ancient aminoacyl–tRNA synthetase gene duplications. Proc Natl Acad Sci USA 1995; 92:2441–2445.
60. Marsh TL, Reich CI, Whitelock RB, Olsen GJ. Transcription factor IID in the Archaea: sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes. Proc Natl Acad Sci USA 1994; 91:4180–4184.
61. Langer D, Hain J, Thuriaux P, Zillig W. Transcription in Archaea: similarity to that in Eucarya. Proc Natl Acad Sci USA 1995; 92:5768–5772.
62. Keeling PJ, Doolittle WF. Archaea: narrowing the gap between prokaryotes and eukaryotes. Proc Natl Acad Sci USA 1995; 92:5761–5764.
63. Koonin EV, Mushegian AR, Galperin MY, Walker DR. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. Mol Microbiol 1997; 25:619–637.
64. Olendzenski L, Hilario E, Gogarten JP. Horizontal gene transfer and fusing lines of descent: the Archaebacteria—a Chimera? In: Syvanen M, Kado C (eds). Horizontal Gene Transfer. London: Chapman and Hall, 1998, pp. 349–362.
65. Gogarten JP. The early evolution of cellular life. Trends Ecol Evol 1995; 10:147–151.
66. Doolittle WF. Phylogenetic classification and the universal tree. Science 1999; 284:2124–2129.
67. Woese CR, Olsen GJ, Ibba M, Soll D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 2000; 64:202–236.
68. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene

clusters. Genetics 1996; 143:1843–1860.
69. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Ann Rev Microbiol 2000; 54:641–679.
70. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7. Nature 2001; 409: 529–533.
71. Dykhuizen DE, Green L. Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 1991; 173:7257–7268.
72. Feil EJ, Holmes EC, Bessen DE, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci USA 2001; 98:182–187.
73. Hilario E, Gogarten JP. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. Biosystems 1993; 31:111–119.
74. Ibba M, Bono JL, Rosa PA, Soll D. Archaeal-type lysyl–tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. Proc Natl Acad Sci USA 1997; 94:14383–14388.
75. Gogarten JP, Olendzenski L. Orthologs, paralogs and genome comparisons. Curr Opin Genet Dev 1999; 9:630–636.
76. Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG, Gogarten JP. Horizontal transfer of archaeal genes into the Deinococcaceae: detection by molecular and computer-based approaches. J Mol Evol 2000; 51:587–599.
77. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 1999; 399:323–329.
78. Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 1996;273:1058–1073.
79. Deckert G, Warren PV, Gaasterland T, et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 1998; 392:353–358.
80. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 1999; 96:3801–3806.
81. Zhaxybayeva O, Gogarten J. Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. BMC Genomics 2002; 3:4.
82. Nesbø CL, Boucher Y, Doolittle WF. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol 2001; 53:340–350.
83. Bellemare G, Vigne R, Jordan BR. Interaction between *Escherichia coli* ribosomal proteins and 5S RNA molecules: recognition of prokaryotic 5S RNAs and rejection of eukaryotic 5S RNAs. Biochimie 1973; 55:29–35.
84. Nomura M, Traub P, Bechmann H. Hybrid 30S ribosomal particles reconstituted from components of different bacterial origins. Nature 1968; 219:793–799.
85. Wrede P, Erdmann VA. Activities of *B. stearothermophilus* 50S ribosomes reconstituted with prokaryotic and eukaryotic 5S RNA. FEBS Lett 1973; 33:315–319.
86. Daya-Grosjean L, Geisser M, Stoffler G, Garret RA. Heterologous protein-RNA interactions in bacterial ribosomes. FEBS Lett 1973; 37:17–20.
87. Asai T, Zaporojets D, Squires C, Squires CL. An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria [see comments]. Proc Natl Acad Sci USA 1999; 96:1971–1976.
88. Wang Y, Zhang Z, Ramanan N. The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. J Bacteriol 1997; 179:3270–3276.
89. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content [see comments]. Nat Genet 1999; 21:108–110.
90. Fitz-Gibbon ST, House CH. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 1999; 27:4218–4222.
91. Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: impli-

cations for comparing genomes on different levels. Genome Res 2000; 10:808–818.
92. House CH, Fitz-Gibbon ST. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J Mol Evol 2002; 54:539–547.
93. Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. Genome Res 1999; 9:550–557.
94. Olendzenski L, Zhaxybayeva O, Gogarten JP. What's in a tree? Does horizontal gene transfer determine microbial taxonomy? In: Seckbach J (ed). Symbiosis. Dordrecht, The Netherlands: Kluwer, 2002, pp. 63–78.
95. Nesbø CL, L'Haridon S, Stetter KO, Doolittle WF. Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between Archaea and Bacteria. Mol Biol Evol 2001; 18:362–375.
96. Darwin C. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray, 1859.
97. Futuyma DJ. Evolutionary Biology. Sunderland, MA: Sinauer Associates, 1986.
98. Hennig W. Phylogenetic Systematics. Urbana: University of Illinois Press, 1966.
99. Barns SM, Delwiche CF, Palmer JD, Pace NR. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci USA 1996; 93:9188–9193.
100. Wheeler DL, Chappey C, Lash AE, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2000; 28:10–14.

10
细菌基因组如何变化

Timothy D. Read and Garry S. A. Myers

引言

微生物基因组天生就是动态的。经过从几代到无数代的繁衍，它们获得突变并进行基因缺失、摄取（acquisition）及重排。某些基因组突变得以保存并在后代中固定下来，成为新的进化潜在路径。20 世纪后 25 年，微生物分子遗传学的发展以及近期比较基因学的出现，已提供了大量基因组变化的知识。尽管要花费大量人力和财力去获得高质量的全序列，但组成基因组序列 A、T、G、C 的排列只不过代表漫长岁月中的某一瞬间，这并不会降低基因组序列对微生物学深入研究的价值。然而，如果要掌握基因组动力学机制，还需要从序列分析中获取更多有用信息。

本章讨论了影响细菌基因组的各种因素，以便了解进化过程的本质。目前，尽管许多研究都以病原微生物为例（反映了整个基因组序列目前的基金偏向），但它们确实给出了微生物基因组动力学的普遍规律，显然，有趣的是，有些变化过程并不是发生在所有基因组上。

选择、漂移及基因组改变

在谈到微生物基因组变化时，需要注意选择（selection）在观察研究中所起的作用。有些突变也许真是中性的，对生物总的适应性没有影响，另一方面，如果非同义突变（nonsynonymous mutation）发生在单拷贝基因，而该基因负责编码的蛋白质对细胞功能是必需的，将会形成致死表型，因此，不会生存下来并被检测到。同理，检测到的增加或减少适应性突变，会以比它们实际发生的频率更高或更低。随机遗传漂移（在一个种群中，某突变频率的增加或减少完全是由于偶尔因素所造成的）是另一个在检测基因组改变时所必须考虑的因素。漂移经常在小群体中起主要作用，例如，导致有害突变在通过种群瓶颈效应后被固定下来。在几乎没有水平基因转移（lateral gene transfer, LGT）的条件或机会下，隔离种群，如昆虫的内共生菌巴克纳氏菌（*Buchnera*）[1]和 *Wigglesworthia glossinidia*[2]直接受随机遗传漂移的影响，这些微生物基因组减少到了极限[3]。

突变可能包括比点突变范围更大的变化，但是这些突变仍然需要用选择或漂移的观点来考虑。许多人类病原体的荚膜基因是高度可变的（由近缘基因组的研究得出；见参考文献［4］和［5］）也是一个需要仔细考虑选择作用的例子。这些位点是高频缺失和重新导入囊膜基因的热点，可能因为它们两侧有多个插入序列（IS）。另一方面，对逃避寄主免疫的荚膜结构的选择，可能提高了荚膜基因位点所表现出的变化频率。

在大肠杆菌（*Escherichia coli*）通过“致病适应性突变（pathoadaptive mutation）”向致病志贺氏菌属（*Shigella*）进化的过程中，表现了受选择压力调控大片段基因组的丢失[6]。赖氨酸脱羧酶（由 *cadA* 编码）的产物 1,5-戊二胺，被认为是一种抗毒力因子，它阻断了志贺氏菌质粒编码肠毒素的活性（由水平转移获得）。许多导致 *cadA* 基因失活的改变，包括插入序列、噬菌体插入及其他基因的重排，已在多种志贺氏菌中发现[6]。这些致病适应性突变的趋同进化显示，生活环境（这里指寄主组织）选择了那些由于 *cadA* 基因功能丢失（即肠毒素活性增强）而使毒力和适应力增加的克隆。

这些例子说明，基因组的改变通常可被多种途径所影响，这一章将探讨一些知道得比较清楚的基因组改变途径。

点突变

点突变是基因组改变的驱动力，通常由损伤核苷酸不精确修复所引起。以处于对数生长期的大肠杆菌（*E.coli*）细胞为例，自发点突变率通常为每碱基每世代 5×10^{-10}[7]。尽管早期模式认为点突变均匀分布，但可能由于差异突变（differential mutation）或选择，突变在整个基因组的分布有某些倾向性。

其中一种倾向性是转换突变（A⇔G 或 C⇔T）高于颠换突变的频率，这种增加是由于修复过程保护受损核苷酸的环状结构所引起。非常相似基因组间的比较表明，转换比颠换高出 2～4 倍（表 1）。另一种一致报道的倾向性为鸟嘌呤和胞嘧啶（G+C）失衡，即嘌呤核苷酸在前导复制链上占优势[8]。G+C 失衡［与操纵子失衡一样，即某些特定核苷酸“串（word）”的不均匀分布[9]］；是一种预测基因组复制原点和复制终止区的有力工具。其他倾向性，包括离复制原点稍远的点突变频率高[10]、在可读框中增加密码子第三位点的突变频率[11]和减少高度表达基因发生突变的可能性[12]。

表 1　近缘基因组间的转换和颠换比较[a]

	转换			颠换				
	T⇔C	A⇔G	总数	C⇔A	C⇔G	T⇔A	T⇔G	总数
肺炎衣原体[b] (*Chlamydophila pneumoniae*)	137	138	275	16	9	9	18	52
布鲁氏杆菌[c] (*Brucella spp.*)	128	152	280	52	64	29	52	197
蚜虫巴克纳氏菌[d] (*Buchnera aphidicola*)	283	230	513	118	35	420	136	709
肺炎链球菌[e] (*Streptococcus pneumoniae*)	7068	7025	14093	1603	725	1527	1652	5552
幽门螺杆菌[f] (*Helicobacter pylori*)	25721	25566	51287	4295	2895	2729	4298	14217

[a]使用 MUMmer 计算（MUMsize = 20；http://www.tigr.org/software/mummer/），只前导链；[b]肺炎衣原体 AR39 对肺炎衣原体 CWL029；[c]猪布鲁氏杆菌 1330 对马尔他布鲁氏杆菌 16M；[d]蚜虫巴克纳氏菌 Ap 对蚜虫巴克纳氏菌 Sg；[e]肺炎链球菌 R6 对肺炎链球菌 TIGR4；[f]幽门螺杆菌 J99 对幽门螺杆菌 26695。

以往认为，基因组的突变率已经进化到了尽可能低的程度[13]，许多能够阻止突变的细菌系统［例如脱氧核糖核酸（DNA）修复］的存在也证明了这一点[14]。细菌 DNA 修复系统的特异性和效率各不相同，因此，不同基因组可能发生不同形式的变化。在这方面，意识到突变本身可以影响突变频率是很有趣的。

例如，某些病原菌在适应新环境的过程中，它们的 DNA 错配修复系统获得了缺陷[15]。在自然分离株中经常观察到有高突变率的亚群，如囊肿性纤维化肺感染中的绿脓假单细胞菌（*Pseudomonas aeruginosa*）[16]，由于 DNA 错配的错误倾向性修复，这些突变株有很高的突变频率。对致病大肠杆菌，当需要快速适应寄主新的生态环境时，这也为高频突变的转变提供了选择优势[17]。在生物体遭遇一段时间的诱变适应后，将会恢复有效的错配修复，与完整基因的重组是一种恢复机制，这解释了观察到的嵌合基因的嵌合现象，以及为什么革兰氏阴性菌修复基因的种系发生重建与典型管家基因的种系树不一致[18]。

外部环境是影响细菌基因组序列改变性质和速率的一个重要因素，然而，在大多数情况下，对这些“外在”影响的研究比对内在详细机制的研究少得多。当考虑环境影响的时候，值得注意的是，在微生物遗传学一些早期研究中，认为诱变剂，如紫外线和亚硝基胍能提高大肠杆菌 DNA 损伤和突变率。众所周知，能引起高速率 DNA 损伤的环境，如高温（热球菌属 *pyrococcus*）、紫外线和干燥（异常球菌属 *Deinococcus*），使生活在其中的细菌获得比大肠杆菌更高效修复 DNA 损伤和断裂的能力。然而，即使是已被深入研究的细菌，对常见的复杂生态位（如土壤和人类小肠）以及通常所接触的温度和 pH 所造成的环境损伤效果，至今仍不十分清楚。在考虑环境变异潜力时，也要考虑 DNA 修复系统中的基因表达效果，例如，有氧和缺氧环境中错配修复系统基因表达是否处在同样的速率呢？

由同源重组、滑动和异常重组引起的插入、缺失和转变

插入和缺失（当不考虑变化过程时统称作插入/缺失［*indels*］）是基因组序列改变的重要途径，基因组中的重复序列是插入/缺失发生的重要焦点。细菌重复序列通常分为两类：低复杂性重复序列（串联重复）和较长重复序列。低复杂性重复序列由小寡核苷酸（从 1 个到 5 个核苷酸）组成，其头尾相连重复许多次。较长重复序列则包括转座元件、大串联重复和间隔重复。已有许多机制解释串联重复产生，包括滑动链错配、同源重组不对等交换，滚动环以及重插入环切（circle excision with reinsertion）[19]。核苷酸组成的倾向性已经证明对串联重复形成频率有很大的影响[20]。

曾认为原核生物高度压缩的基因组中不存在大串联重复，与真核生物中的复制机制类似，大串联重复有可能在原核生物中通过利用低复杂性重复序列作为引物而产生[20,21]。另外，大重复序列在抗原改变中起作用，并可组成某些基因组的大部分片段，如支原体（*Mycoplasma*）的简并基因组（reduced genome）。在生殖道支原体（*Mycoplasma genitalium*）中，编码黏附蛋白 MgPa 的一个三基因操纵子的重复竟占该基因组的 4%以上[22]。这些重复序列相似性为 78%～90%，相似到足以发生同源重组的程度，但又不会到一旦重组就发生整个基因转变的程度。另一具有生物学重要意义的大串

联重复的例子是肺炎衣原体（*Chlamydophila pneumoniae*）的 *tyrP* 基因，它编码了一个重要芳香氨基酸转运蛋白。从世界各地的分离株都含有 1～3 个该基因的串联重复，而 2～3 个重复与呼吸道分离株有关，有一个重复单位则与心血管分离株相关（图 1；R.Belland，私人通讯，2003）。低复杂性重复序列，又被称为可变数目串联重复（variable numbers of tandem repeats，VNTR），已经运用到细菌流行病学和法医鉴定中，如炭疽芽孢杆菌（*Bacillus anthracis*）不同菌株中核苷酸变化很小[23, 24]。事实上，多聚核苷酸串（以一个核苷酸为重复单位）为最常见的 VNTR，通常是细菌基因组中最具变化性的特征之一。已发现炭疽芽孢杆菌（*B.anthracis*）Ames 基因组中的一个 35 聚体核苷酸的 VNTR，以每代 10^{-4}的频率增加或减少 1 个核苷酸单位（P.Keim，私人通讯，2003）。重复单位越多，重复拷贝数的变化越小。

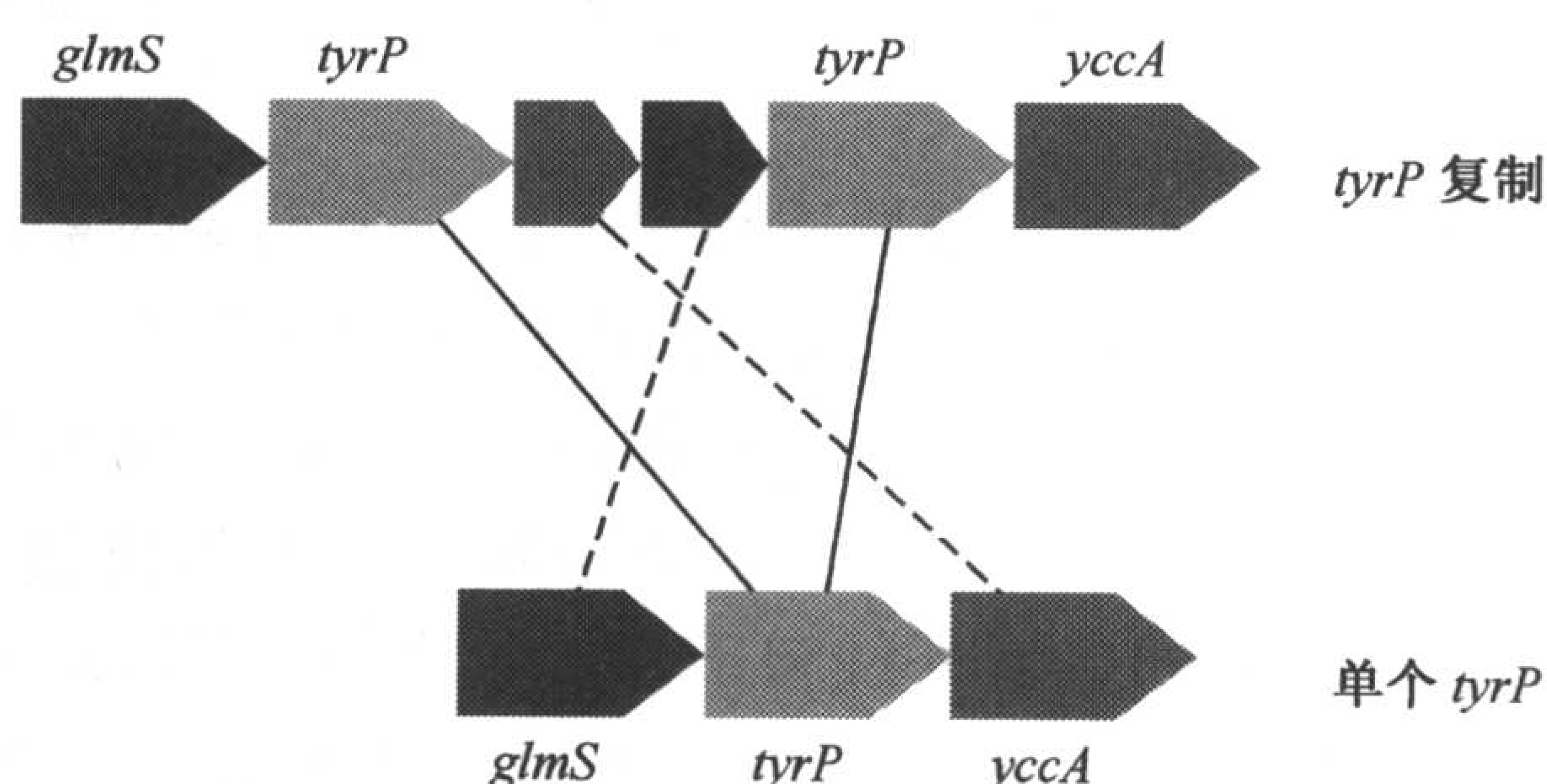


图 1　肺炎衣原体（*C. pneumoniae*）*tyrP* 基因座位的重复和疾病类型有关。*TyrP* 多拷贝基因型与呼吸系统感染有关，而单拷贝基因型与心血管状况关系密切；*glmS*，*yccA* 的残留和 *tyrP* 中单核苷酸多态性分布，揭示这种单拷贝基因型更古老。*glmS*，葡糖胺-果糖-6-磷酸氨基转移酶基因；*tyrP*，酪氨酸/色氨酸特异转运蛋白基因；*yccA*，膜整合蛋白（可能通透酶）（资料由 R. Belland 友情提供）。

在某些情况下，VNTR 会影响基因的功能，例如，在流感嗜血菌（*Haemophilus influenzae*）基因组中，很多基因都在 5′端包含短 VNTR，而这些 VNTR 中核苷酸重复单位的长度不是 3 的倍数。通过高度可变序列来调控基因的这种相似模式，也在病原菌，如空肠弯杆菌（*Campylobacter jejuni*）[25]、脑膜炎奈瑟氏球菌（*Neisseria meningitides*）[26]和肺炎链球菌（*Streptococcus pneumoniae*）[5]中发现。VNTR 单位改变可以导致移码突变，这种突变要么使某基因失去功能，要么使曾经发生过移码突变的某基因重新正常表达。在另外一样流感嗜血菌（*Haemophilus influenzae*）相变异（phase variation）例子中，纤毛结构基因的表达是由 Poly-AT VNTR 隔开的双向启动子控制[27]。如果 VNTR 含 9 个单位，那么启动子之间的距离使纤毛基因有效表达，如果是其他单位数，细菌将不产生纤毛。

比较基因组研究发现，短正向序列（有时短至 7 个碱基）通常含有 1000 个以内核苷酸序列的插入/缺失作用位点[28, 29]。然而，一般认为，由于重复序列长度太短，无法通过交联桥的形成和 Holliday 结构的分离来完成由 RecA 介导的同源重组。正向重复序列间距离越远，重复序列越短，由 RecA 介导同源重组的频率就越低。对这种短重复序列间的异常重组，提出了两种机制：第一它是由 DNA 复制中后随链错误配对（链滑动

错配)；第二与外切核酸酶产生的同源小片段配对[30]。

基因转换（gene conversion）是由同源重组中同源序列的不对等交换而产生，基因转换是酵母基因组中接合型转变的机制[31]。基因转换使细菌基因组上的多基因家族协同进化（在较长时间内保持基因间高水平相似性）[32,33]。在淋病奈瑟氏球菌（*Neisseria gonorrhoeae*）中，表达菌毛亚单位基因和高达 19 份分布在不同位点无义拷贝的重组，导致菌体表面结构的抗原性变化[34]。对该菌毛基因中间体（intermediate pilus gene）序列的详细分析表明，基因转换涉及同源重组和短同源序列间不依赖 RecA 的序列交换[34]。

染色体复制和终止区的倒位

复制起始位点和终止区域对称重组代表发生在整个基因组范围内的改变（图 2），这个现象首先在肠杆菌科中发现，随着比较基因组测序的深入，发现在细菌中广泛存在[35,36]。似乎合理的解释是双向复制叉发生错误导致双链断裂，从而促进了距起始位点等距离 DNA 之间的重组[37]，因为在复制终止处更可能产生 DNA 断裂结构，而在终止位点区域对称倒位的频率比起始位点区域高[36,38]。

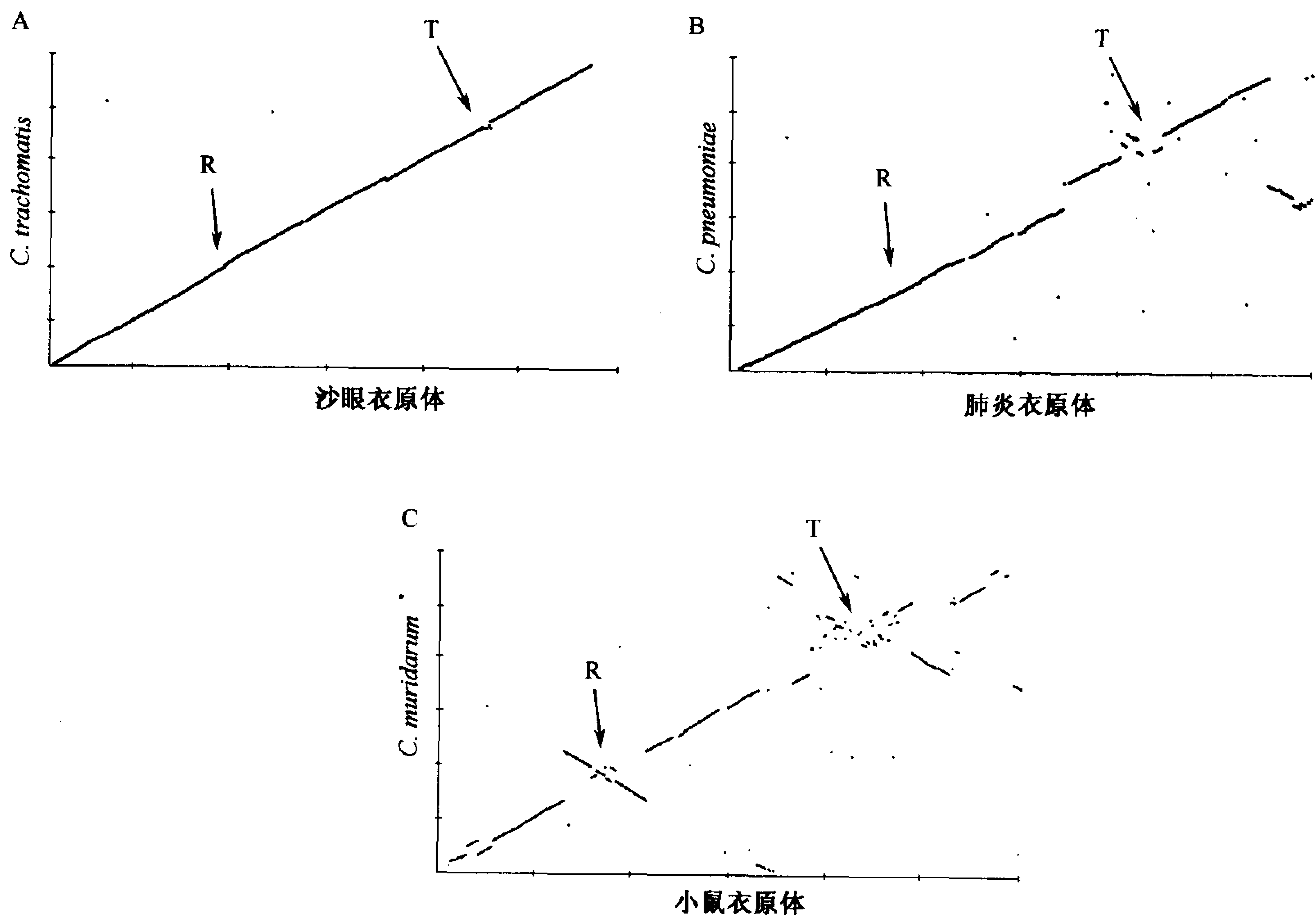


图 2（I） 通过衣原体科的对比显示了在复制起点和终点附近染色体同线性的破坏。每一个小点分别代表相似性最高的蛋白质。R 和 T 分别表示复制起点和终点。A. 沙眼衣原体（*Chlamydia trachomatis*）对小鼠衣原体（*Chlamydia muridarum*）. B. 肺炎衣原体（*Chlamydia pneumoniae*）对豚鼠衣原体（*Chlamydia caviae*）. C. 小鼠衣原体（*Chlamydia muridarum*）对豚鼠衣原体（*Chlamydia caviae*）（经许可从参考文献［38］复制）。

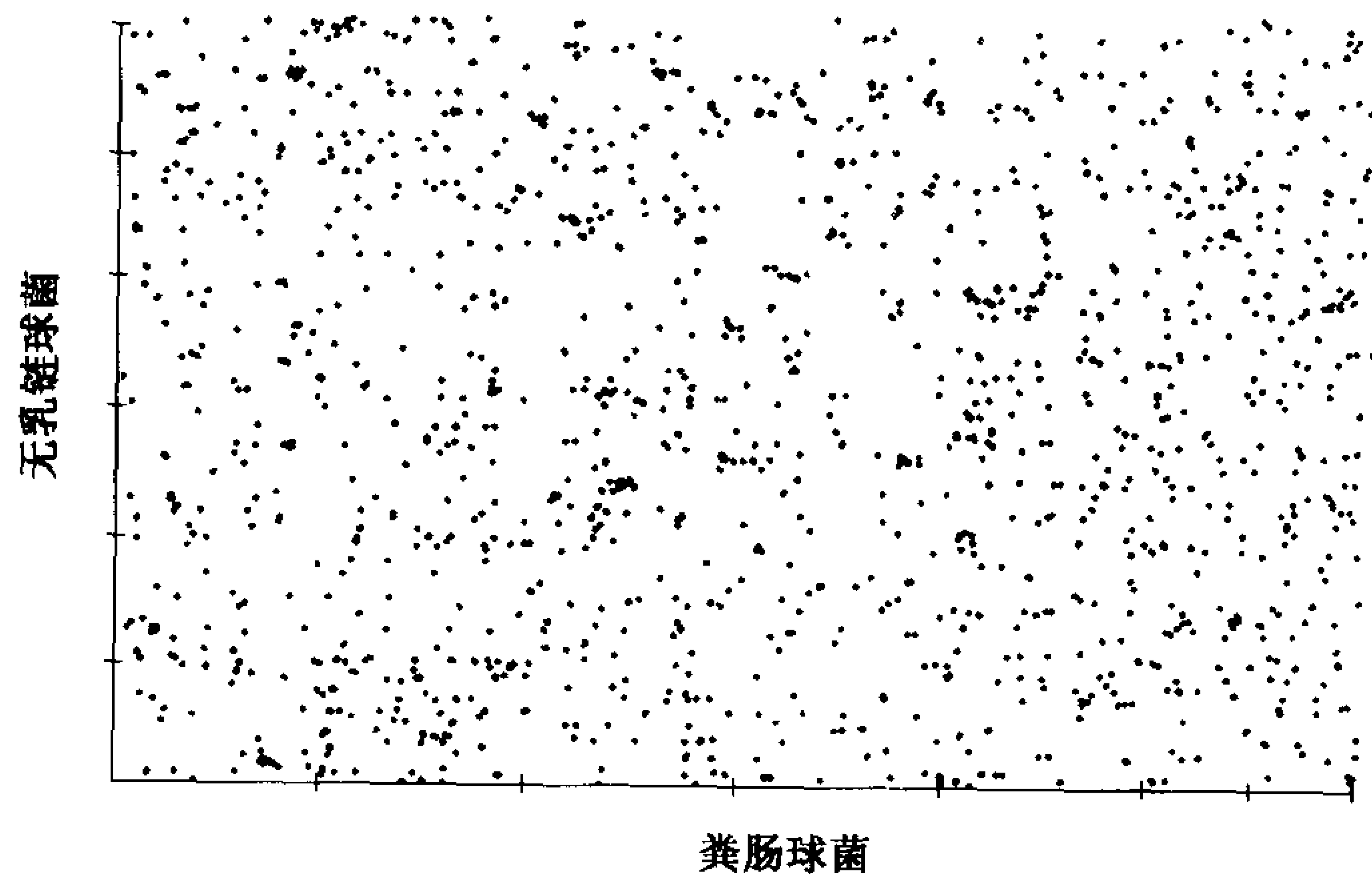


图 2（Ⅱ） 粪肠球菌（*Enterococcus faecalis*）与无乳链球菌（*Streptococcus agalaciae*）的比较显示它们没有大量染色体同线性。

比较基因组学研究揭示基因组的另一个特征是，功能保守基因分布在起始位点附近，而在终止位点附近含有大量的假定基因和可能通过种间横向获得的基因[38~41]。有几种理论解释这种现象：在终止位点处有更高突变率和较低表达水平，或高频率外源DNA的插入。如果复制终止位点处突变率较高，管家基因（housekeeping gene）则不应在这里分布，管家基因在复制原点和非必需基因在终止区的不平衡分布，可能进一步加剧突变在基因组中固定的倾向性。复制原点处保守基因的大量分布，可导致对复制引起对称性倒位的选择性限制，因此，这类变化在复制起始区比终止区较少出现，该例子说明，由于多种因素的影响，细菌基因组动力学可能非常复杂。

将两个近缘基因组作图时，并不是所有细菌都存在典型X形模式，例如，粪肠球菌（*Enterococcus faecalis*）和无乳链球菌（*Streptococcus agalactiae*）（图2），几乎没有保守基因序列，可能是粪肠球菌中发生过大量重组[42]，因为该菌含有许多插入元件、原噬菌体和整合质粒。

程序化的基因组改变

淋病奈瑟氏球菌（*N. gonorrhoeae*）的菌毛通过基因转变造成抗原性改变[34]代表一类高频率转换，这具有依照程序进行改变的特征。程序化改变不是由于DNA修复和复制错误，或“自私”DNA元件活动所引发的单一事件，这些变化使其能够应对频繁发生的环境变化，通常赋予细菌一种选择优势。

许多确定的程序化改变包括位点特异性重组系统，位点特异性重组酶在特异的靶位点催化DNA交换，在许多经充分研究的系统中，重组导致基因组中相近两个或多个靶位点的倒置。

肺炎链球菌（*S. pneumoniae*）Ⅰ型限制系统特异性亚基的调节是一个例子[5]，在*hsdS*、*hsdM*和*hsdR*基因（分别编码三亚基酶的特异性、甲基化和限制性亚基）旁边，有一个位点特异性重组酶，它可导致*hsdS*基因和邻近的*hadS*假基因中两个位点的倒

置（图 3）。该系统用随机鸟枪法对几个序列克隆测序时发现的，这样给出了 4 种核苷酸序列。据推测，这些可变基因可为限制酶提供不同特异性靶位点，这是一种防止病毒感染的适应机制。在肺炎支原体（*Mycoplasma pulmonis*）中也发现了类似的系统[43,44]。众所周知的其他例子包括沙门氏菌（*Salmonella hin*）鞭毛的相变系统，摩拉氏菌（*Moraxella piv*）[45]菌毛的可变系统以及类细菌（*Bacteroides*）多个表面蛋白可变系统[46]。

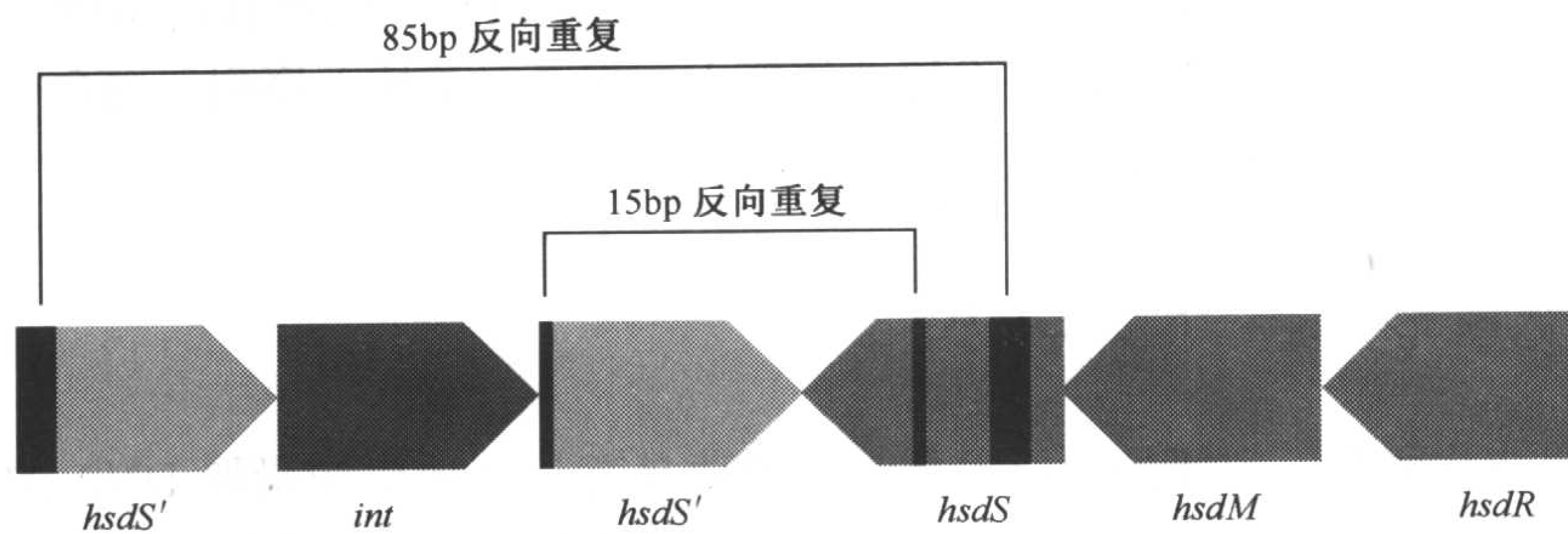


图 3　肺炎链球菌（*C. Pneumoniae*）TIGR4 多态 I 型限制酶（*hsdS*）操纵子的结构。标为 *hsdS'* 基因是部分 *hsdS* 基因（假基因），*int* 是一个整合酶基因，*hsdSMR* 分别是特异性，修饰性和限制性亚单位。反向 85bp 与 15bp 插入重复序列作了标记，以 85bp 或 15bp 反向重复序列为边界，观察到将 *hsdS* 基因与 *hsdS'* 假基因融合在一起的克隆。

随着测序研究对基因组理解的深入，程序化改变将会普遍发现。其他例子，包括由滑动链错配导致的基因型和蛋白质表达[47]，可能是利用自私元件来控制细菌的某些生物特性。例如，在枯草芽孢杆菌（*Bacillus subtilis*）中，*skin* 前噬菌体的切除重建了 *sigK* 基因，这是孢子形成的关键[48]。

可移动遗传元件和基因组重排

基因组中可移动遗传元件是基因组发生内在变化的重要原因。细菌基因组中的插入序列（IS）是最常见的可移动元件，一个基本的插入序列元件包括一个位点专一性重组酶（转座酶）和侧翼 DNA 序列，它们经常与重组位点的反向重复序列相连[49]，侧翼通常有短正向重复序列，这些重复序列来自转座期间产生的短单链 DNA 末端。

插入序列元件的重要特点，是在基因组内部不同位点之间转座（跳跃），它能以一个或多个单位的转座子移动，在某些情况下，侧翼的反向重复序列可移动插入框，可通过反式作用转座[50]。

几种细菌的基因组含有多个相同插入序列，这种扩增是通过非保守转座来实现[42,51]。在这种情况下，多个插入序列元件变成同源重组位点，引起诸如缺失、倒位和插入这种基因的重排。在不同插入序列元件家族之间，转座机制和靶 DNA 的专一性位点有极大差异，例如，IS7 和 IS30，对某特定序列的插入都有非常明显的专一性[52,53]。另一方面，IS1 似乎对插入位点的腺嘌呤核苷酸和胸腺嘧啶核苷酸有一定偏爱性[54]，通过避免插入到对寄主生存至关重要的序列中，插入元件的位点专一性是一种与特定寄主长期共存的适应，例如，在肺炎链球菌（*S. pneumoniae*）基因组 84 个插

入序列元件中，仅有2个插入到功能基因中[5]。

研究较少的其他可移动元件，包括内含子（intron）、反转子（retron）和内含肽（intein）。细菌含有Ⅰ型和Ⅱ型内含子，它们可从转录RNA分子上剪切下来[55]，而内含肽元件可从翻译的蛋白质上剪切下来，反转子是通过编码反转录酶实现自身复制的小遗传元件[56]。许多细菌基因组含有由重复序列组成的复合结构[57,58]，如肺炎链球菌（*S. pneumoniae*）[5]的BOX元件，奈瑟氏球菌属（*Neisseria*）的Correia元件以及大肠杆菌（*E. coli*）的REP元件[59]。尽管重复序列经常分散在不同菌株的不同基因组位点中，显示出具有一定的能在基因组内移动的能力，但是对于插入序列和内含子这些序列的保存和扩增机制仍不很清楚。

通过水平转移摄取外源DNA

基因组变化的另一主要外在原因是某些细菌具有从其他生物摄取DNA的能力，即所谓水平基因转移（LGT）[60]。通过LGT摄取外源DNA有三种机制：转化、噬菌体介导的转移（转导）和接合转移（图4）。转化是指细菌从环境中获取DNA的过程。1944年，Averye及同事在有关肺炎链球菌（*S. pneumoniae*）的研究中发现，DNA是细胞内的遗传物质[61]。某些种类的细菌含有一些"感受态（competence）"基因，来帮助束缚、吸收一些通常具备特异序列特征的DNA。在芽孢杆菌属（*Bacillus*）、链球菌属（*Streptococcus*）、奈瑟氏球菌属（*Neissera*）和嗜血菌属（*Haemophilus*）中都发现了这种感受态（competence）系统[62,63]。外源DNA被吸收后，能够通过重组方式整合到基因组中去，这些菌具很高的重组频率[64~67]。这并非巧合，即使没有特殊DNA吸收基因，某些细菌仍有能力吸收环境中的DNA，例如，Mandel与Higa发现，大肠杆菌（*E. coli*）在氯化钙的环境中具转化能力，这对早期分子克隆技术的发展具有很重要的意义[68]。

可自主复制和移动的质粒在无亲缘关系细菌间的移动过程，称为接合（conjugation）。以F质粒为例，肠杆菌科（*Enterobacteriaceae*）提供了便于理解的模型[69]，F质粒的接合转移始于双链DNA分子中的缺刻链，随后，滚动复制产生先导链，单链DNA从供体细胞通过由质粒编码Ⅳ型分泌装置形成连接桥，转移到不含质粒的受体细胞。质粒编码的另一种蛋白质PilE，帮助稳定细菌间的相互接触，从而提高了DNA的转移率。一旦转到受体细胞中，线型单链DNA环化，并能复制重新构成原来的双链环状分子。有一种较小"可移动（mobilizable）"质粒，虽然缺乏自身的接合基因，但是含有关键的DNA靶序列，使它们能利用同一细胞中较大质粒编码的蛋白实现转移[69]。

大量研究表明，结合性DNA可以在种系关系很远的生物间转移[70]，最令人吃惊的是，竟然能发生在大肠杆菌（*E. coli*）和酵母（yeast）之间。大多数接合性质粒不能稳定存在于广泛的寄主系统中，这一特性会导致"自杀性"DNA转移，即质粒复制子在接合后，从受体中很快丢失，但是，有些短片段可通过转座或重组过程整合到新基因组。更有甚者，有些质粒还能进行反向转移（retrotransfer），刚刚抵达受体细胞的这些质粒会通过接合转移反向进入供体细胞[71]。这样，从受体染色体转移到迁徙质粒的DNA序列，最终可抵达最初的供体细胞，含有这种质粒的细菌，如同基因"扒手"一

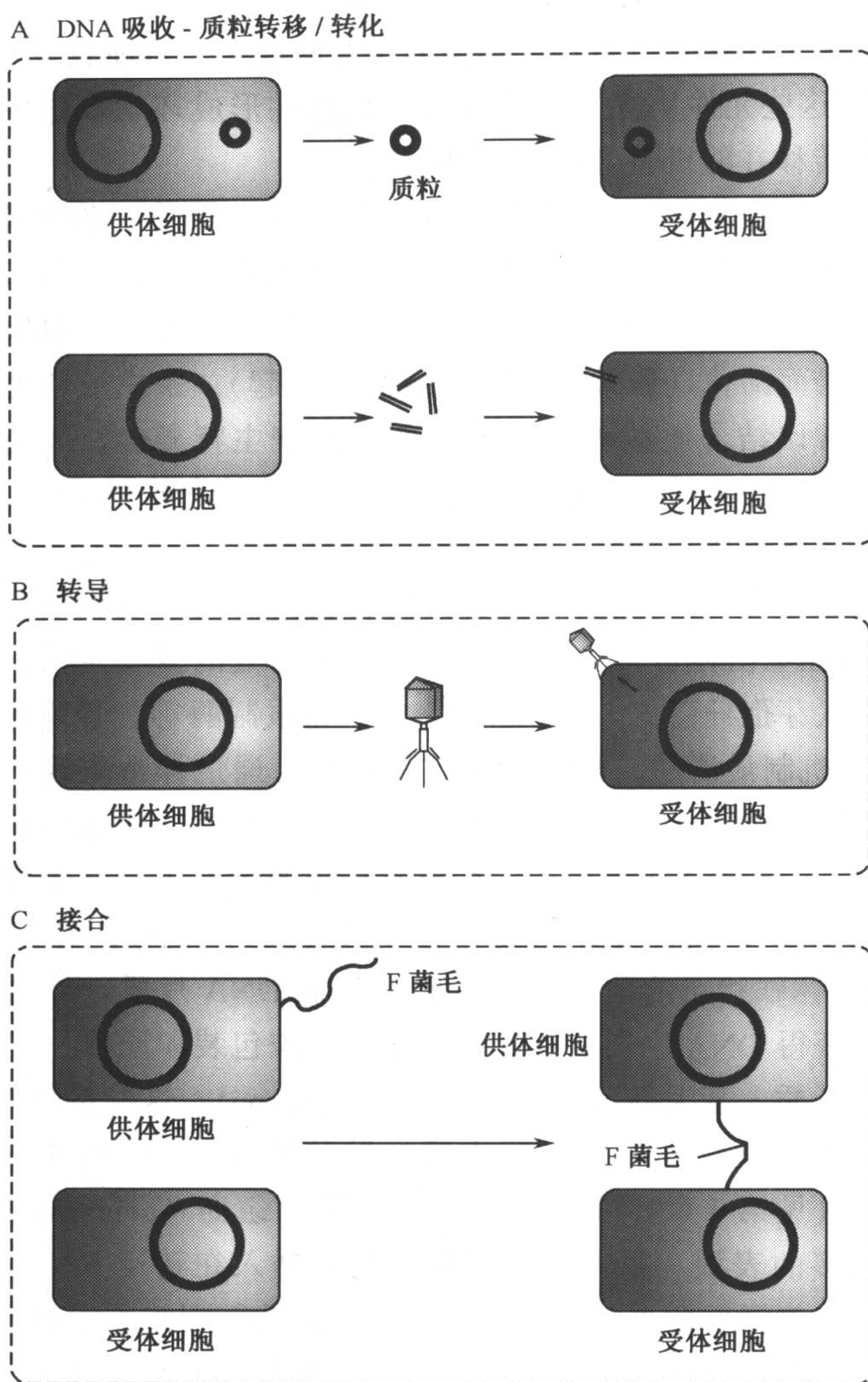


图 4 通过水平基因转移（LGT）获取外源 DNA 的机制

样，有效地利用反向转移获得它们临近细菌的基因。

另一深入研究的接合系统是根癌土壤杆菌（*Agrobacterium tumefaciens*）中的 Ti 质粒，它能通过与 F 质粒相似的结构，将一段自主移动的 DNA 片段转移到植物基因组中[72]。大多数情况下，在肠杆菌科（*Enterobacteriaceae*）外发现的质粒是能够自主移动的，但它们很少或几乎不含类似 F 质粒的接合决定因子。例如，肠球菌的信息素诱导质粒编码一系列基因产物，使寄主细胞能够感知无质粒细胞所编码（外源染色体编码）和分泌的信息素[73]，在对信息素的应答中，激活接合作用，将质粒转移到无质粒细胞中，质粒的其他功能却防止细胞自身信息素对自身应答的诱导。因此，这种形式的 DNA 移动是由胞间信号分子精密调控的一种复杂细菌行为。

有许多这样的质粒家族，每种质粒对不同信息素作出反应，从而使细菌的性状能迅速传播给其他肠球菌。最近基因组研究表明，由信息素诱导的质粒和接合转座子是（细

菌）获得万古霉素（vancomycin）抗性的关键因素[42]。

与噬菌体、转座子和接合质粒特征类似的接合转座子，目前只发现在革兰氏阳性细菌基因组中[74,75]，这些元件从寄主菌基因组中切除并形成环状双链中间体，随后通过滚环方式进行接合转移并随机整合到受体基因组。

噬菌体以另一种途径将外源 DNA 整合到细菌基因组[60,75]。许多噬菌体可将自身整合进细菌基因组（溶源现象），而不是感染后立即杀死寄主，在整合过程中，前噬菌体抑制自身大部分基因的转录而进入潜伏状态。当它们从基因组中切除并进入裂解循环时，它们就会去阻遏而活化（通常由寄主的紊乱所引起），寄主通常死亡并产生大量病毒子代。噬菌体通常比结合质粒表现出更为狭窄的寄主特异性，这是由几个方面决定的：黏附所需细菌表面特异受体、重组酶介导与基因组整合的靶位点，以及噬菌体与细菌基因调控系统的协同进化。

噬菌体的切除和整合（以及质粒的获得和丢失），体现了近缘细菌间基因总含量的重要区别[76]。溶源性噬菌体常常含有能给寄主带来好处的重要基因，如毒素（在霍乱菌 *Cholera*[77]和梭状芽孢杆菌 *Clostridium*[78]中）和限制酶等。噬菌体还用于使基因组 DNA 迁移，转导的机制最清楚，前噬菌体在切除时，通过异常重组可以把与其相邻的寄主基因组 DNA 连带切除并一起包装进病毒的衣壳，随后，成熟噬菌体的感染使非噬菌体 DNA 通过同源重组整合到新寄主。在缺乏转导的情况下，感染噬菌体如果含有与寄主高度同一性的 DNA，就能重组进基因组 DNA。必须记住，在细菌基因组中位于 *att* 位点之间并处于休眠状态的前噬菌体，能像其他 DNA 片段一样通过转座或插入来获得基因，这些新获得 DNA 序列能在噬菌体复苏时被包装和迁移。

比较测序发现了插入基因岛（islands of inserted gene），这些基因作为一个整体进行插入和移动。最初被称为致病岛[79]，因为第一批发现的这些区域含有病原性肠道细菌株的毒性必需基因（见第 4 章），现在更准确地称基因组岛（genomic island），因为这些区域所含基因有更多的表型。第一批发现的岛均是插入在转运 RNA 基因（tRNA）附近，含整合酶基因和侧翼反向重复序列。然而，最近的基因组测序显示出更复杂和有趣的情况，即许多岛具有噬菌体、转座子、接合质粒和接合转座子特征[75,80]。

在考虑外源 DNA 如何改变细菌基因组时，阻碍基因摄取效率的屏障也应值得重视。这些主要是生态学的许多生物，如胞内菌和极端环境微生物，在它们生存的环境中，只有非常有限的其他供体和受体生物。关键基因的少量表达或不表达，同样会限制如噬菌体和 IS 插入序列元件的作用。正如上面所讨论的，复制基因在其他寄主中不表达，限制了许多细菌质粒的寄主范围。“混杂”质粒，如 IncP 非兼容群，有很广的寄主范围，通过特殊性的适应性它们克服了上述有限寄主范围问题[81]。寄主细胞错配修复的特异性，是在亲缘关系很远的细菌间进行同源重组的一道屏障。在重组链形成的异源二倍体中，错配碱基越多，重组效率越低，正如大肠杆菌（*E.coli*）/沙门氏菌（*Salmonella*）之间的重组那样[82]。因此，许多由重组介导的机制，局限于种内或高度保守的 DNA 区域中。由 Lawrence 等提出寄生 DNA 摄取的另一道屏障，是在重复序列间重组时存在丢失的偏向性[83]，这能解释在许多生物体中假基因被较快地删除，以及细菌基因组为什么不能持速增大。

尽管其他因素也看似合理，但是限制/修饰系统通常是细菌对噬菌体和质粒入侵的

一种防御策略[84]。这种防御机制的潜在问题是：如果细菌只有单个限制或修饰系统，逃避甲基化的噬菌体能导致整个细菌群落的消亡。许多质粒和噬菌体有逃避限制性作用的系统[85]，脑膜炎奈瑟氏球菌（*N. meningitidis*）[86]的基因组序列揭示了两种机制来对付这种情况。首先，在基因组中发现了20种以上的限制/修饰系统；其次，其中有6种系统与通过滑动链错配产生相变化的重复序列有关[84]，这两种机制确保了脑膜炎奈瑟氏球菌（*N. meningitidis*）种群的限制/修饰系统的多样性。

在其他细菌中重复序列也与限制/修饰基因的调控有关，包括流感嗜血菌（*H. influenzae*）Rd[87]中甲基转移酶的相变化调控，以及限制/修饰 *hsd* 基因的倒置元件调控（图3），这些重复序列也可能参与长距离的同源重组。此外，限制/修饰系统不仅在灭活噬菌体DNA方面发挥作用，而且还能把外源DNA有用片段加工成足够长的片段，以有利于重组，并能剔除其周围的有害片段。

外源DNA序列的整合，其本身可能是基因组改变的另一种选择。如果一个外源基因可为细胞带来有用的表型（如抗生素抗性），就可能有一个改良的过程。如果外源DNA发生了突变，而突变体的密码子使用偏好正好与寄主自身可读框的密码子使用偏好相似，那么该突变体就可能保留。其他改良新获得DNA的压力，可能是为了避免限制位点或是根据这些序列与复制叉相对位置而调整其嘌呤含量。在大多数情况下，外源DNA不会编码对寄主有用的功能而保留，因此将积累点突变和突变而逐渐消亡。尽管IS元件不见得会编码有用功能或通过在基因组内的转座而复制，但它们却能够在新寄主基因组中存活较长时间。因此，尽管最初携带的质粒、噬菌体或转化DNA早已从基因组中丢失，IS元件的存在也许是古代LGT变化的遗迹。

整合子：程序化的外源基因转移

细菌是否已进化成如下的机制：从侵入DNA中获取特定序列，或让质粒从寄主染色体中获取基因？这可能是整合子基因捕获系统存在的目的[89]。整合子通常在质粒和转座子的重组盒上发现，整合子的主要组成部分，包括整合酶基因，位点特异重组靶位点（通常是59bp的序列）以及使基因在重组盒中表达的启动子。紧邻59bp序列的基因，能被整合酶介导的环化过程（circulation）和整合所“捕获”，整合酶也可反向发挥作用，即从整合子上切除基因盒。因此，自然发生的整合子由高度可变的基因复合体组成，其中最好的例子是在霍乱弧菌（*Vibrio cholerae*）基因组中发现的巨大质粒[90,91]，最近的证据表明了整合子/基因盒具有普遍性和多样性，它们可能代表着一个巨大的可变基因组模板，有助于LGT和细菌基因组的进化[92]。

第四维基因组学

比较测序的新时代将给静态单个基因组序列以时间为尺度提供新的一维领域，并加深对细菌进化动力学的理解。关键不但要找出相应的机制，还要给出影响整个基因组变化的相对速率。基因组动力学特性将反映出生物的生活习性及其所面对的选择压力。对基因组变化不同速率的更深入理解，最终能导致重建更复杂的种系发生模型。

早期比较基因组学在专性胞内寄生菌，如衣原体（*Chlamydia*）、立克次氏体（*Rickettsia*）以及共栖菌，如奈瑟氏球菌属（*Neisseria*）和肠球菌属（*Enterococcus*）中，发现了两种不同模式。专性胞内寄生病原体是由自由生活的祖先通过简并方式进化而来的，基因由于缺失和缺乏获取新序列的机会，重复序列（包括 IS 元件）逐渐从它们的基因组中丢失[93]。衣原体（*Chlamydia*）有最少的和稳定的基因含量，插入、倒位及缺失几乎都不可避免地产生灾难性后果。

Suyama 和 Bork 提出[36]，胞内病原菌核苷变化的频率，似乎比其他菌复制原点的依赖性倒位的频率高。一种解释是细菌 DNA 修复系统对复制叉处双链断裂的修复更有效。很少证据表明胞内寄生菌能吸收外源 DNA，这是由于它们有优越的胞内生态位和缺乏 LGT 机制，但这不是绝对的，衣原体的复制终止区可能是个重组热点，即所谓“可塑性区域”，不同衣原体可通过这一区域进行基因交换。

相反，共栖菌的生存是产生了不同的基因组动力学。这些生物在特定而多样的微生物环境中生存并互相竞争，而且还要躲避寄主免疫系统不断变化的应答反应。一般情况下，这些菌有更多重复序列、IS 元件和位点特异重组系统。相对于胞内寄生菌，共栖菌基因的倒位和缺失频率比点突变大得多，新近获得了很多这样的基因（表现为噬菌体、质粒和致病岛的形式）。

影响基因组改变的机制、速率和因素等方面将是另一研究领域，也将对新兴的以基因组为基础的微生物法医学产生深远影响。现在能在几天内得到微生物基因组序列草图，将来还可能开发出更快速、经济的 Sanger 测序反应和其他技术（如以微阵排列为基础的再测序），通过这些技术可轻松地获得来自犯罪案例或流行病调查中多菌株的完整基因组序列数据。第一个实例是 2001 年美国生物恐怖袭击中的一株炭疽芽孢杆菌（*B. anthracis*）序列草图的获得[29]，将生物恐怖中的 Ames 菌株序列与实验室中的 Ames 菌株染色体进行比较，显示了 11 处相同单核苷酸多态性。结果证明这些多态现象为实验室菌株所特有，它是通过化学诱变和高温处理来稳定细胞中的毒性质粒 pXO1 和 pXO2。

涉及多个近缘基因组的研究将会鉴定出菌株谱系特征性标记，并能进一步绘出细菌的“家谱（pedigree）”，这种有关突变速率的知识，对理解这些发现的意义是必需的。例如，如果对两个基因组测序，发现它们每一个核苷酸都是相同的，这意味着什么？要使它们在进化上产生分枝最多需多少代？在以相对较高的速率积累突变的基因组区域内发现了变化吗？这些发现的变化也可能提供有关菌株最近发生历史的线索。在实验室生长了数代的菌株，可能积累了移码或非同义的突变，而这些突变基因可能在自然分离株中是必须基因（例如，炭疽芽孢杆菌（*B. anthracis*）的实验室 Ames 菌株，在编码整体调控蛋白（global regulator）的两个基因 *relA* 和 *spo0A* 中有非同义突变[29]）。不同类型变化（单核苷酸置换、串联重复、转座等）的相对速率，也能反映菌株的生长环境。

随着基因组大量涌入公共数据库，将会更容易地认识到基因组的动态性。利用这些数据来建立基因组变化的精确模型，这会对未来的生物学产生巨大影响。

（吕颂雅　译）

参考文献

1. Tamas I, Klasson L, Canback B, et al. Fifty million years of genomic stasis in endosymbiotic bacteria. Science 2002; 296:2376–2379.
2. Akman L, Yamashita A, Watanabe H, et al. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat Genet 2002; 32:402–407.
3. Wernegreen JJ. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet 2002; 3:850–861.
4. Tettelin H, Masignani V, Cieslewicz MJ, et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci USA 2002; 99:12391–12396.
5. Tettelin H, Nelson KE, Paulsen IT, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 2001; 293:498–506.
6. Day WA Jr, Fernandez RE, Maurelli AT. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. Infect Immun 2001; 69:7471–7480.
7. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics 1998; 148:1667–1686.
8. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 1996; 13:660–665.
9. Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF. Skewed oligomers and origins of replication. Gene 1998; 217:57–67.
10. Mira A, Ochman H. Gene location and bacterial sequence divergence. Mol Biol Evol 2002; 19: 1350–1358.
11. Wright F, Bibb MJ. Codon usage in the G+C-rich *Streptomyces* genome. Gene 1992; 113:55–65.
12. Sharp PM, Li WH. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for “rare” codons. Nucleic Acids Res 1986; 14:7737–7749.
13. Kimura M. On the evolutionary adjustment of spontaneous mutation rates. Genet Res, 1967; 23–34.
14. Giraud A, Radman M, Matic I, Taddei F. The rise and fall of mutator bacteria. Curr Opin Microbiol 2001; 4:582–585.
15. Taddei F, Matic I, Godelle B, Radman M. To be a mutator, or how pathogenic and commensal bacteria can evolve rapidly. Trends Microbiol 1997; 5:427–428; discussion 428–429.
16. Oliver A, Canton R, Campo P, Baquero F, Blazquez J. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. Science 2000; 288:1251–1254.
17. Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. Role of mutator alleles in adaptive evolution. Nature 1997; 387:700–702.
18. Denamur E, Lecointre G, Darlu P, et al. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell 2000; 103:711–721.
19. Romero D, Palacios, R. Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet 1997; 31:91–111.
20. Achaz G, Rocha EP, Netter P, Coissac E. Origin and fate of repeats in bacteria. Nucleic Acids Res 2002; 30:2987–2994.
21. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 1987; 4:203–221.
22. Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995; 270:397–403.
23. Schupp JM, Klevytska AM, Zinser G, Price LB, Keim P. *vrrB*, a hypervariable open reading frame in *Bacillus anthracis*. J Bacteriol 2000; 182:3989–3997.

24. Keim P, Price LB, Klevitska AM, et al. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J Bacteriol 2000; 182:2928–2936.
25. Parkhill J, Wren BW, Mungall K, et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 2000; 403:665–668.
26. Saunders NJ, Jeffries AC, Peden JF, et al. Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. Mol Microbiol 2000; 37:207–215.
27. van Ham SM, van Alphen L, Mooi FR, van Putten JP. The fimbrial gene cluster of *Haemophilus influenzae* type b. Mol Microbiol 1994; 13:673–684.
28. Michel B. Illegitimate recombination in bacteria. In: Charlebois R (ed). Organization of the Prokaryotic Genome. Washington, DC: ASM, 1999, pp. 129–150.
29. Read TD, Salzberg SL, Pop, M, et al. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 2002; 296:2028–2033.
30. Conley EC, Saunders VA, Saunders JR. Deletion and rearrangement of plasmid DNA during transformation of *Escherichia coli* with linear plasmid molecules. Nucleic Acids Res 1986; 14: 8905–8917.
31. Haber JE. Mating-type gene switching in *Saccharomyces cerevisiae*. Annu Rev Genet 1998; 32: 561–599.
32. Pride DT, Blaser MJ. Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. J Mol Biol 2002; 316:629–642.
33. Hashimoto JG, Stevenson BS, Schmidt TM. Rates and consequences of recombination between rRNA operons. J Bacteriol 2003; 185:966–972.
34. Howell-Adams B, Seifert HS. Molecular models accounting for the gene conversion reactions mediating gonococcal pilin antigenic variation. Mol Microbiol 2000; 37:1146–1158.
35. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 2000; 1:1–9.
36. Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 2001; 17:10–13.
37. Tillier ER, Collins RA. Genome rearrangement by replication-directed translocation. Nat Genet 2000; 26:195–197.
38. Read TD, Myers GS, Brunham RC, et al. Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. Nucleic Acids Res 2003; 31:2134–2147.
39. Read TD, Peterson SN, Tourasse N, et al. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 2003; 432:81–86.
40. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc Natl Acad Sci USA 2000; 97:14668–14673.
41. Lecompte O, Ripp R, Puzos-Barbe V, et al. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res 2001; 11:981–993.
42. Paulsen IT, Banerjei L, Myers GS, et al. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science 2003; 299:2071–2074.
43. Gumulak-Smith J, Teachman A, Tu AH, Simecka JW, Lindsey JR, Dybvig K. Variations in the surface proteins and restriction enzyme systems of *Mycoplasma pulmonis* in the respiratory tract of infected rats. Mol Microbiol 2001; 40:1037–1044.
44. Sitaraman R, Denison AM, Dybvig K. A unique, bifunctional site-specific DNA recombinase from *Mycoplasma pulmonis*. Mol Microbiol 2002; 46:1033–1040.
45. Tobiason DM, Lenich AG, Glasgow AC. Multiple DNA binding activities of the novel site-specific recombinase, Piv, from *Moraxella lacunata*. J Biol Chem 1999; 274:9698–9706.
46. Krinos CM, Coyne MJ, Weinacht KG, Tzianalbos AO, Kasper DL, Conestack LE. Extensive surface diversity of a commensal microorganism by multiple DNA inversions. Nature 2001;

414:555–558.
47. Hood DW, Deadman ME, Jennings MP, et al. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc Natl Acad Sci USA 1996; 93:11,121–11,125.
48. Sato T, Harada K, Kobayashi Y. Analysis of suppressor mutations of *spoIVCA* mutations: occurrence of DNA rearrangement in the absence of site-specific DNA recombinase SpoIVCA in *Bacillus subtilis*. J Bacteriol 1996; 178:3380–3383.
49. Chandler M, Mahillon J. Insertion sequence revisited. In: Lambowitz AM (ed). Mobile DNA II. Washington, DC: ASM, 2002, pp. 305–366.
50. Chen Y, Braathen P, Leonard C, Mahillon J. MIC231, a naturally occurring mobile insertion cassette from *Bacillus cereus*. Mol Microbiol 1999; 32:657–668.
51. Lawrence JG, Ochman H, Hartl DL. The evolution of insertion sequences within enteric bacteria. Genetics 1992; 131:9–20.
52. Peters JE, Craig NL. Tn7: smarter than we thought. Nat Rev Mol Cell Biol 2001; 2:806–814.
53. Olasz F, Kiss J, Konig P, Buzas Z, Stalder R, Arber W. Target specificity of insertion element IS30. Mol Microbiol 1998; 28:691–704.
54. Zerbib D, Gamas P, Chandler M, Prentki B, Bass S, Galas D. Specificity of insertion of IS1. J Mol Biol 1985; 185:517–524.
55. Belfort M, Reaban ME, Coetzee T, Dalgaard JZ. Prokaryotic introns and inteins: a panoply of form and function. J Bacteriol 1995; 177:3897–3903.
56. Lampson B, Inouye M, Inouye S. The msDNAs of bacteria. Prog Nucleic Acid Res Mol Biol 2001; 67:65–91.
57. Liu SV, Saunders NJ, Jeffries A, Rest RF. Genome analysis and strain comparison of correia repeats and correia repeat-enclosed elements in pathogenic *Neisseria*. J Bacteriol 2002; 184: 6163–6173.
58. Correia FF, Inouye S, Inouye M. A 26-base-pair repetitive sequence specific for *Neisseria gonorrhoeae* and *Neisseria meningitidis* genomic DNA. J Bacteriol 1986; 167:1009–1015.
59. Meyer BJ, Schottel JL. Characterization of *cat* messenger RNA decay suggests that turnover occurs by endonucleolytic cleavage in a 3' to 5' direction. Mol Microbiol 6, 1095–1104 (1992).
60. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299–304.
61. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types—induction of transformation by a deoxyribonucleic-acid fraction isolated from pneumococcus type-III. J Exp Med 1944; 79:137–158.
62. Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev 1994; 58:563–602.
63. Dubnau D. DNA uptake in bacteria. Annu Rev Microbiol 1999; 53:217–244.
64. Vazquez JA, Berron S, O'Rourke M, et al. Interspecies recombination in nature: a meningococcus that has acquired a gonococcal PIB porin. Mol Microbiol 1995; 15:1001–1007.
65. Suerbaum S, Achtman M. Evolution of *Helicobacter pylori*: the role of recombination. Trends Microbiol 1999; 7:182–184.
66. Feil EJ, Maiden MC, Achtman M, Spratt BG. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol Biol Evol 1999; 16: 1496–1502.
67. Meats E, Feil EJ, Stringer S, et al. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J Clin Microbiol 2003; 41:1623–1636.
68. Mandel M, Higa A. Calcium-dependent bacteriophage DNA infection. J Mol Biol 1970; 53: 159–162.
69. Zechner EL, de la Cruz F, Eisenbrandt R, et al. Conjugative-DNA transfer processes. In: Thomas CM (ed). The Horizonal Gene Pool. Amsterdam: Harwood, 2000, pp. 87–174.

70. Bates S, Cashmore AM, Wilkins BM. IncP plasmids are unusually effective in mediating conjugation of *Escherichia coli* and *Saccharomyces cerevisiae*: involvement of the *tra2* mating system. J Bacteriol 1998; 180:6538–6543.
71. Szpirer C, Top E, Couturier M, Mergeay M. Retrotransfer or gene capture: a feature of conjugative plasmids, with ecological and evolutionary significance. Microbiology 1999; 145:3321–3329.
72. Christie PJ, Vogel JP. Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. Trends Microbiol 2000; 8:354–360.
73. Clewell DB. Bacterial sex pheromone-induced plasmid transfer. Cell 1993; 73:9–12.
74. Salyers AA, Shoemaker NB, Stevens AM, Li LY. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. Microbial Rev 1995; 59:579–590.
75. Osborn AM, Boltner D. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid 2002; 48:202–212.
76. Banks DJ, Beres SB, Musser JM. The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. Trends Microbiol 2002; 10:515–521.
77. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 1996; 272:1910–1914.
78. Fujii N, Oguma K, Yokosawa N, Kimura K, Tsuzuki K. Characterization of bacteriophage nucleic acids obtained from *Clostridium botulinum* types C and D. Appl Environ Microbiol 1988; 54:69–73.
79. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 1997; 23:1089–1097.
80. Osborn AM, da Silva Tatley FM, Steyn LM, Pickup RW, Saunders JR. Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. Microbiology 2000; 146:2267–2275.
81. Perlin MH. The subcellular entities a.k.a. plasmids. In: Yasbin RE (ed). Modern Microbial Genetics. New York: Wiley-Liss, 2002, pp. 507–560.
82. Matic I, Rayssiguier C, Radman M. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell 1995; 80:507–515.
83. Lawrence JG, Hendrix RW, Casjens S. Where are the pseudogenes in bacterial genomes? Trends Microbiol 2001; 9:535–540.
84. Blumenthal RM, Cheng X. Restriction-modification systems. In: Yasbin RE (ed). Modern Microbial Genetics. New York: Wiley-Liss, 2002, pp. 177–225.
85. Wilkins BM. Plasmid promiscuity: meeting the challenge of DNA immigration control. Environ Microbiol 2002; 4:495–500.
86. Tettelin H, Saunders NJ, Heidelberg J, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
87. De Bolle X, Bayliss CD, Field D, et al. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol 2000; 35:211–222.
88. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 1997; 44:383–397.
89. Hall RM, Collis CM. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol Microbiol 1995; 15:593–600.
90. Clark CA, Purins L, Kaewrakon P, Manning PA. VCR repetitive sequence elements in the *Vibrio cholerae* chromosome constitute a mega-integron. Mol Microbiol 1997; 26:1137–1138.
91. Heidelberg JF, Eisen JA, Nelson WC, et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 2000; 406:477–483.
92. Holmes AJ, Gillings MR, Nield BS, Mabbutt BC, Nevalainen KM, Stokes HW. The gene cassette metagenome is a basic resource for bacterial genome evolution. Environ Microbiol 2003; 5:383–394.
93. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends Genet 2001; 17:589–596.

11
基因组学时代的细菌生物多样性概念

Frederick M. Cohan

引言

毋庸置疑，如果要对细菌的生态多样性做出全面评价，就需要用到以基因组和序列为基础的技术方法。由于目前只有一小部分细菌是可培养的，所以要全面确定细菌的生物多样性，最好的方法是从细菌所处的自然生境中扩增基因并进行序列分析[1,2]。此外，以基因组和序列为基础的技术方法，还可以揭示极其相近（已培养）物种之间的生态多样性。研究表明，已经定种的、在生态学上各自独特的类群，如果不考虑它们之间的生态学特征，可能属于相同的序列簇[3~5]，根据其基因组特征也可将这些类群归为一个簇[6,7]。除了寻找生态学上独特的种群之外，基因组学方法还能帮助阐明每个群体成员的生态学功能[8,9]，以及不同种群在环境中是如何分布和共存的。我们还能用此方法研究水平转移，是如何使细菌在新的生态小生境[10]中占优势，此外，还可鉴定水平转移物质的供体菌[11]、基因转移的遗传学及生态学屏障[12]。

为了达到以上这些目标，基因组学正在进行着一场革命。鉴定生态学上独特的菌株，并确定哪些菌株在生态学上是可互换的，进而探询同一种群中各个成员的组成情况，这些工作是十分重要的。例如，将来可以深入研究水平转移在细菌入侵新生态小生境中所起的作用。生态学上两个独特种群的基因水平差异可以表现为，是与种群间生态学差异相关水平转移物质的差异。从另一个角度看，能区分同一种群在生态学上可互换菌株的水平转移物质，可视为是水平转移这颗陨星在染色体上留下无生态意义的“弹痕”[13]。本章综述了最新的一些用以揭示并对细菌多样性进行分类的研究方法，为对基因组特征和表达进行研究提供良好的生态学基础。

首先，要了解细菌分类学在将菌株分为生态学独特类群中所起的指导性意义。在细菌分类学水平，通常是根据表现型、遗传学及生态学上的独特性，将细菌划分为不同类群，在这些类群之间存在着巨大的缺口[14~16]。与其他生物相同，将划分的细菌类群进行认证并命名为种。细菌的种最初总是根据表现型确定，判断的理论基础是该类群是否具有某种代谢能力（见第 9 章）[4,16]。如今，随着分子生物学技术的出现，可以将这些根据表现型分析所划分的相似类群进一步校正、合并、定种。例如，目前全基因组脱氧核糖核酸（DNA）—DNA 杂交法，已作为区分不同细菌种的基本准则。如果某一细菌类群的基因组 DNA 与另一类群的染色体 DNA 之间的复性率达到 70%或以上，并且与以往用表现型分类所得的结果相对应，那么我们就可以将这两个类群归为同一个种[17,18]。最近，16S 核糖体 RNA（rRNA）序列相似性也用于鉴定种，其原理是如果 16S rRNA 的序列差异达到 2.5%以上，就视为不同种（尽管有些不同种间的 16S rRNA 序列差异可能小于这一标准）[19]。

这些通过表型特征、全基因组和序列相似性定名的细菌种，是否与生态学上独特种群、与其他生物界（动、植物界）中已划分的种相一致呢？如果希望通过研究序列和基因组数据，来收集与微生物群落生物多样性及其生态学功能相关的信息，这些问题至关重要。

在高等有性真核生物界中，种具有相同的动力学特性，这种特性使相同种内的生物具有高度同质性，而种间则具有异质性[20]。种具有一个很基本的特性，即同一种中的遗传多样性是由一种凝聚力限制着[4,21,22]。以高等真核生物为例，基因交换能力认为是多样性变异的主要阻力[23,24]；种的第二个特点是不同物种间的分离不可逆，一旦物种到达变异的交歧点，该物种就能不受拘束地向其他方向发生变异[25,26]，在高等有性真核生物界中，这个变异关口很可能是繁殖方式的分歧[23]；最后，当同一种中不同成员在生态学上可互换时，通过分享各自适合的资源与条件，不同种间可以达到共存[27]。因此，种的这些动力学性质适用于除动、植物等外，其他具有特殊性别特征的类群上包括细菌。

目前，在微生物进化遗传学家中，已有越来越多的人达成了共识：在细菌分类学中已定名的种，并不存在独特的动力学特征。首先，细菌分类学较多的展示了典型种的代谢及假定的生态多样性[13,28,29]；第二，近几十年来，分类学中 DNA-DNA 杂交研究的应用，在已定名的物种中揭示了多层次遗传多样性，并在几个不同的种中选取两个或两个以上菌株进行基因组测序，已经证实了这些结论的准确性；第三，多位点测序分型（MLST），揭示了在已定名的物种中存在着多基因序列簇[3,37,38]，并且这些序列簇可能与生态学上独特种群相对应[5,37,39]；最后，环境中生态和序列多样性的自然历史学研究证明，存在着多种生态学上的独特类群，它们的序列极其相似，可以将它们划分到同一个已定名的种中[40~43]。

显然细菌界中的生态学多样性，比已有准确数目、已定名的物种数目要多得多，然而，在已定名的一个物种中、每个生态学上的独特类群中是否有可归于同一种的动力学特性，这个问题还存在着争议。本章节阐述了一些有关生物多样性实质的、引人注目的最新观点：生物种概念在细菌中的应用[39,44,45]、生态型概念[4,46]以及撇开物种的细菌多样性概念[12]。由于这些概念起源于细菌基因交换本质中的一些不同假说，将对细菌基因交换的特质和不同基因交换率在细菌种群动力学产生的后果进行综述。

细菌基因交换的特点

基因交换的稀有性

细菌可以在无限的世代中克隆和复制，而单菌落偶尔可能发生重组，即一个供体菌染色体上的一小段取代了受体菌染色体上同源的一段。我们可以评估出自然界中细菌基因重组的效率，利用“回顾（retrospective）”的方法，在自然群落中寻找序列或等位基因的变异，其基本原则是：当个体中不同位点的突变遗传因子具有高度连锁性（即连锁不平衡）或不同 DNA 片段产生出一致的种系发生树时，可以推测其基因片段具有低重组效率[47]。

这些“回顾”的方法，揭示了大多数细菌种群间发生基因片段的重组概率与突变概

率相当或稍高一些[48~50]。例如，金黄色葡萄球菌是一个严格无性繁殖的细菌分类单元，其基因片段的重组率比突变频率要低 3 倍[51,52]。脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）则是比较容易发生重组的一种细菌，其基因片段的重组率比突变频率要高 3.6 倍[37]。

由于基因重组比单核苷酸突变所涉及的核苷酸要多得多，因此，对一段特定的核苷酸序列，重组的影响比突变要高 80 倍[37]，这一结果支持了这样一种说法：在细菌中重组并不稀少，而依赖稀有重组的细菌进化模式是不合理的[12,45]。然而，用最新以序列为基础的方法对重组率进行评估，所得的每个基因片段发生重组的频率，与用早先以等位基因为基础的方法所得的结果相差无几（即重组发生频率比突变率低 10 倍），正如我们在本章中所讨论的，这种重组稀有性使自然选择在细菌种群内遗传多样性的发展上有十分深远的意义。

基因交换的混杂性

由于细菌基因重组的稀有性，在基因交换过程中对交换对象的选择也就没过分地关注，细菌可与 DNA 序列相差 25％以上的菌发生同源重组[53~56]。

尽管如此，不同细菌种群发生基因交换还存在重要的限制因素，包括要求供体菌与受体菌必须有相同的重组载体（对于噬菌体和质粒介导的重组）和微生境[57,58]。而且，不同供体序列的整合也是同源重组的分子限制因素，重组发生要求供体片段末端与受体同源区域能很好地配对，如果供体与受体的序列存在很大差异，重组就不可能发生[54,59,60]。此外，细菌错配的修复系统，一旦发现供体与受体片段之间核苷酸错配存在，就很有可能对重组整合进行校正[55,61]。在肠杆菌科中[55,61]，错配修复在性别分离中起重要作用，而在革兰氏阳性菌中这种情况较少[53,59]。

细菌重组不仅限于同源片段的转移，在异源重组中细菌可以“捕获”其他细菌的基因位点和操纵子[62]。最近的基因组分析得知，细菌基因组的一部分（一般为 5％～10％）可能来自不同种的细菌[63]。

虽然，基因水平转移在细菌基因组中留下很明显的标记，并不能说水平转移频率比较高，尤其是在每个细胞的水平上。Lawrence[45]评估认为，在大肠杆菌中成功的水平转移 600～700 万年才能发生，一个中等大小为 10^{12}的种群，每个细胞每一世代发生成功水平转移的频率为 7×10^{-22}（假定每小时分裂一代）[64]，尽管水平转移频率低于百万分之一，每个细胞的水平转移率（无论成功与否）可以达到 7×10^{-16}。

稀有单混杂基因交换引发进化的后果

从其他分类单元引入适应性 DNA

细菌基因交换经常发生，这足以导致适应突变遗传因子从一个种转移到另一个种中。例如，抗性突变基因可从奈瑟氏球菌转移到链球菌中，通过同源重组取代了原先的抗生素敏感性基因[65]，而且，经常发生的异源重组也可以将数以百计的基因位点，引入特定的细菌基因组中[63]。种间 DNA 转移无论是突变基因或新操纵子的转移，都是以很低重组频率就可以实现的基因重组方式。按前面提到的计算结果，即使一百万个水平

转移物质中有一个是适应性的，大肠杆菌的基因水平转移每个细胞的重组率为 7×10^{-16}。引入适应性 DNA 仅需要很低的重组频率，这是因为在一个菌系中引入一个适应性基因只需要发生一次重组，一旦得到这个基因，那么自然选择能提升新适应基因型的丰度。此外，如果以此种方式获得的基因，使受体具有适应某种新生境的生存能力，那么这一重组产生的基因型很可能是成功的[12,66]。

重组在中性序列多样性产生中所起的作用

Maiden 及其工作人员开发了以序列信息为基础的 MLST 技术，将菌株划分为无性繁殖复合群（clonal complexe）系统[3]。MLST 的基本原理是对 7 个持家基因进行测序，而这 7 个基因的多样性在适应性方面比较中性，而这些蛋白不随特异生境的变化而变化，可在生态学上的不同种群之间互相交换，因此，MLST 分析数据凭经验观察重组在中性序列多样性产生中所起的作用。

通过 MLST 研究的结果显示，重组影响菌株以序列为基础的系统进化分类，例如，脑膜炎奈瑟氏球菌的种系发生史，随着基因的改变而发生变化[67]，没有单一的生物种系发生史，脑膜炎奈瑟氏球菌每个菌株的基因组，都是已定名的种和外界基因的组合体[12]。

MLST 还表明，重组对中性序列多样性的影响比突变作用要大得多，因为重组发生的频率与每个基因发生突变的频率相当，但每次重组所涉及的核苷酸数目是数以百计。

然而，某一领域的重组并不影响中性序列多样性，这使得采用 MLST 技术将菌株划分为无性繁殖复合群[3]成为可能，该技术中，菌株间的进化距离用不同基因位点进行定量，如果两个菌株在某一基因位点，有一个或多个核苷酸被取代（可能是由重组产生的），就认为它们是不同的。随后，菌株就划分为无性繁殖复合群，在 7 个基因位点中，如果一个菌株与特定中心菌株有 5 个或 5 个以上（有时是 6 个和 6 个以上）的基因位点相同，那么，就可以认为它们属于一个无性繁殖复合群。

随着 MLST 技术的发展，Maiden 及其工作人员发现，用 7 个基因位点划分无性繁殖复合群所得的结果与用 11 个位点的结果一致，而这 7 个基因位点的选择不影响菌株分类结果[3]。虽然，一个特定菌株中有少数基因位点发生重组，这种联合体是划分菌株多样性的有效工具，甚至在迄今为止所知最常发生重组的脑膜炎奈瑟氏球菌中，也能发现明显的无性繁殖复合群。用 MLST 划分的无性繁殖复合群在经验上与生态学上独特的种群相对应，而且，这种对应性有很有力的理论依据。

重组对生态隔离群适应性变异的影响

对于动、植物等高度有性繁殖群体，相近群体间变异的产生需要群体间存在性隔离[23]。假定种群间和种群内的重组率相同，那么，重组就能很快使种群间的差异消失。

相对而言，细菌间的重组率非常低，所以种间重组的发生并不影响种间差异，假设两个基因组在某些位点上存在等位基因差异，使其可利用不同营养物，从而形成在生态上有较大差异的种群，其适应性基因型分别以 ABCD 和 abcd 表示，这种适应性基因型（如 abcde）的重组频率会降到 1-S（S 表示对重组的选择强度）。

以前的研究表明，非适应性外源等位基因的平衡频率（equilibrium frequency）可以

用 C_b/S 表示，C_b 为种群间的重组频率[46]。由于细菌种群内的每个基因的重组频率相当低（$C_w \approx 10^{-6}$，每基因片段），即使细菌种群间具有与种群内同样的重组率，非适应性外源基因出现的频率可忽略不计（细菌基因水平转移的存在与否对生态差异的影响也是如此）。虽然，性隔离在动植物种的进化中起关键作用，但对细菌的进化却无关紧要[4,46]。

重组对细菌种群内多样性的影响

长期以来，无性繁殖群体的遗传多样性，一直认为是通过周期性选择（periodic selection）方式被清除[68]，在没有重组存在的情况下，适应性突变产生适应环境的子代群体，最终会取代其他细胞群体，并导致所有位点上遗传多样性的丧失。

高频率重组毫无疑问可产生细菌的遗传多样性，如果某一适应性突变基因型，通过重组整合到另一遗传背景的基因组中，受体菌的整个基因组得以保留；如果非适应性突变基因型菌株中的一个基因片段，通过重组整合到适应性突变体中，该基因片段也得以保留。周期性选择导致多样性的丧失是对重组的最大挑战，细菌重组的高频率（可达到有性繁殖体水平），才能保持生物多样性[69,70]（图 1）。当周期性选择的强度较高时(如选择强度为 10%)，每一轮选择就可以清洗掉几乎所有的多样性。对于自然界中典型可观察到的重组率（0.3～0.6 倍于突变率），周期性选择摧毁了几乎所有多样性，仅保留 0.001%～0.2%多样性（图 1）；对于中等强度的自然选择（选择强度为 1%），有 0.02%～0.2%的序列多样性保留。因此，即使细菌重组频率达到很高水平，对遗传多样性的贡献也非常有限。

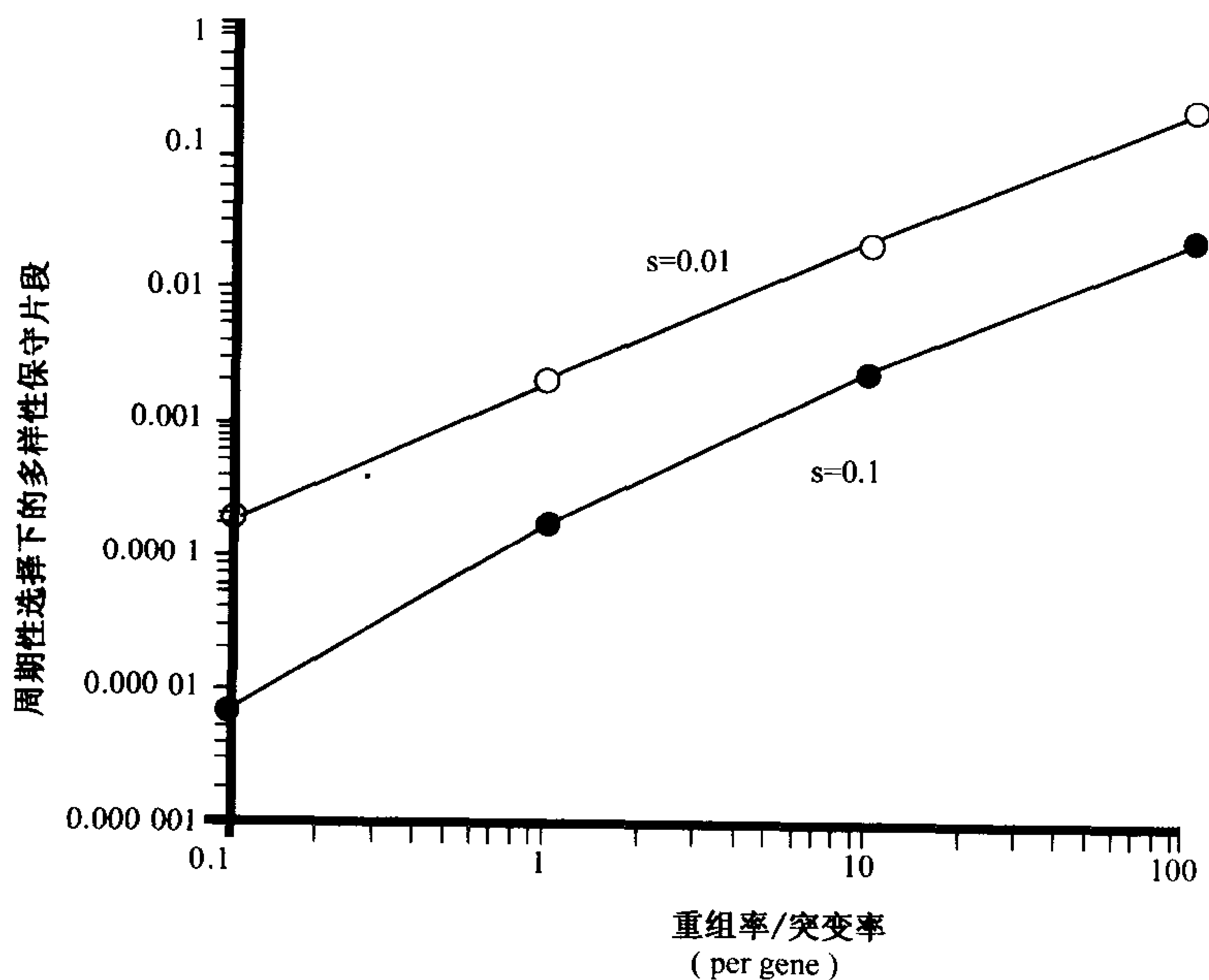


图 1　不同选择条件下重组率和周期性选择对多样性抑制效果的关系。图中所示重组率和突变率的比例，表示在自然界中所能观察到典型的比例范围。纵坐标表示周期性选择后，所保留基因组片段的多样性比例。该结果基于 Monte Carlo 的假设[70]。

周期性选择是否对细菌群体的遗传多样性有如此大的作用？Lawrence[45]对此提出疑问，其理由部分是基于Guttman和Dykhuizen对大肠杆菌基因组序列的研究[71]，他们的研究表明，大肠杆菌基因组大部分具有较高的遗传多样性，但在gapA基因附近有一段约30kb片段的遗传多样性却很低，Lawrence据此认为，周期性选择对细菌整个基因组遗传多样性的影响不大，这与对动物群体的研究结果类似。

Majewski和我先前都认为，Guttman和Dykhuizen关于周期性选择的理论有缺陷，原因是其理论建立在这种假设之上，即在大肠杆菌中，某一种适应性突变体（或重组体）的适应能力大大超过其他突变体[72]。基于对大肠杆菌等种类系统分类学的研究[16]，及其巨大生态多样性有关的进化遗传学的认识[12,73,74]，这种假设是不可能的，因为，每种菌都包括很多适应不同生态环境的菌株，其中某一菌株的适应能力不可能在任何生境中都最强，如前所述，周期性选择对几乎所有细菌的遗传多样性有抑制作用，即使实际的重组频率高于一般观察到的重组频率[69,70]（图1），对他们研究数据的正确解释，需要更多来自对细菌生态多样性的深入研究。

细菌生态学多样性模型

细菌系统分类学研究主要基于对其表型和遗传型的聚类分析[16]，然而，近十年来，细菌进化动力学概念开始影响这一领域。

建立在遗传交换和性隔离基础上的种概念

Mayr对生物意义上“种”这一概念的贡献在于，把生物进化理念带进了系统分类学[23,24]，生物意义上“种”的概念改变了动物学家和植物学家关于对物种到底应如何描述的观念：某一物种不仅是相近有机体的聚类，更应该是在生态和进化上具有特定动力学特性的基本单元。根据Mayr关于物种的概念，物种应该由特定的一类群体组成，在其个体间通过重组抑制变异[21~23]。

Dykhuizen和Green[39]把生物种的概念应用到细菌分类中，他们用种系发生方法，把细菌分成不同单元（簇），每一簇实际代表一个“种”，簇内不同成员在基因上有一定差异，不同簇内的基因存在较大差异。

这种分类方法由于没有把细菌内混交式（promiscuity）基因交换考虑在内，因而一直受到批评[4,12,46]，细菌在簇内和簇间都存在基因交换[12,65,75~77]，种系发生学研究表明，细菌在簇内发生的重组要比簇间发生的重组高得多，条件是不把供体菌划分在系统内，如果把供体菌算进来，那么种间等位基因的水平转移现象就更加常见。

最近提出关于细菌多样性的两个概念，能解释细菌性隔离的进化起源，Lawrence[45]的“没有种的物种形成（speciation without species）”模型，首先提出了由于获得基因的不同，而产生两个生态差异的种群概念，在种群适应性变异过程中，任何种群内基因交换都会受高强度选择。这些基因附近的染色体区域不会形成均一化交换区，因此，中性序列变异在这一区域累积，导致这一区域成功重组的效率下降[44,54]。由于每个种群与其环境有高度适应性，基因的改变遍及整个染色体，每一改变都对其附近区域起保护作用，最终整个染色体被保护起来以避免发生重组，或者说序列多样性阻

止了整个染色体基因交换的发生。

Lawrence 模型阐明了适应性变异基因，导致其侧翼区域遗传交换降低，并由此积累中性序列变异，我对这一点深信不疑。然而，还不清楚这些区域对基因交换的抑制作用，是否为不同生态类型的序列多样性所必须。因为种群间中性序列多样性产生，本质上是自我加速的过程，在这一过程中，任何随机中性序列多样性（基因组中的任何部位，不管是不是被保护的部位）都趋于性隔离的增加，导致重组降低，从而进一步增加序列多样性。早先的研究结果表明，芽孢杆菌属中序列多样性和性隔离之间的正反馈现象，最终导致不同生态类群之间的无限中性序列多样性[78]。无论如何，Lawrence[45]认为，整个基因组中环境适应性基因的积累会加速这一过程，我非常赞同这一点。

“物种形成分子机制”的概念认为，错配修复系统可促进微生境适应性入侵[44,45]。因为有些细菌（如肠杆菌科）[55]的错配修复是导致性隔离的主要因素，有缺陷错配修复系统，增加了其他细菌适应性基因的转移概率。然而，一旦种群适应了新环境，由于产生较高的突变率，错配修复系统的缺陷会逐渐暴露[79]。但通过水平转移，可使该种群重新获得功能性错配修复系统[80]，物种开始由于缺乏完整的功能性错配修复系统，而轻易地获得序列多样性，当错配修复系统重新恢复时，获得的序列多样性就可以导致性隔离[44]。

用生物物种概念可更好理解由性隔离导致进化的动力模型，Dykhuizen 和 Green[39]设计用重组存在与否对物种进行分类，即种类的确定可根据其是否存在高频率重组，根据不同种群间性隔离的累计程度可以理解它们的起源。然而，正如我曾经提出的，不同菌种间的重组并不能阻止其多样性[12,46]，对两个群体而言，一旦其获得具有生态差异的性状，重组率（由于太低）不足以影响适应性群体的整体性，而且，重组不会阻隘环境适应性累积。在细菌中，性隔离进化与永久变异性进化不相关，因此，在细菌系统学中，生物物种概念不合适[4]。

尽管可以准确预测性隔离加快的过程，Lawrence[45]和 Vulic[44]的模型对细菌物种起源这一概念的理解是不必要的。下面介绍的是细菌进化问题，即种群生态多样性在细菌进化中的重要作用。

周期性选择的细菌种概念

我曾给对应于适应性突变（或重组）细菌的“生态型（ecotype）”下过定义：即生态型是指一类具有相同或相似生态位的菌株，同一生态型中适应性突变株的适应能力远超过其他菌株并使后者趋于灭绝，然而，适应性变异并不能引起其他生态位菌株的灭绝[4,46,81]（图 2)。因此，生态型指的是一组菌株，其多样性受周期性选择的制约，而这种选择作用对适应性突变株有利。因为这种选择作用可周期性地把遗传多样性调整到接近零，因此，它对生态型内菌株有非常强的凝聚力。

当两个菌群出现生态上明显差异时，它们就各自经历各自的周期性选择，在这种情况下，有利于适应性突变体的自然选择，只对突变群体中的多样性起抑制作用，这一点对新种形成至关重要。这些变异不可逆转，因为周期性选择不能阻止进一步变异，也不能重组，我曾解释过这一点[46]。

细菌生态型，从周期性选择角度看，具有“种”的特性：一个拥有坚实凝聚力的群

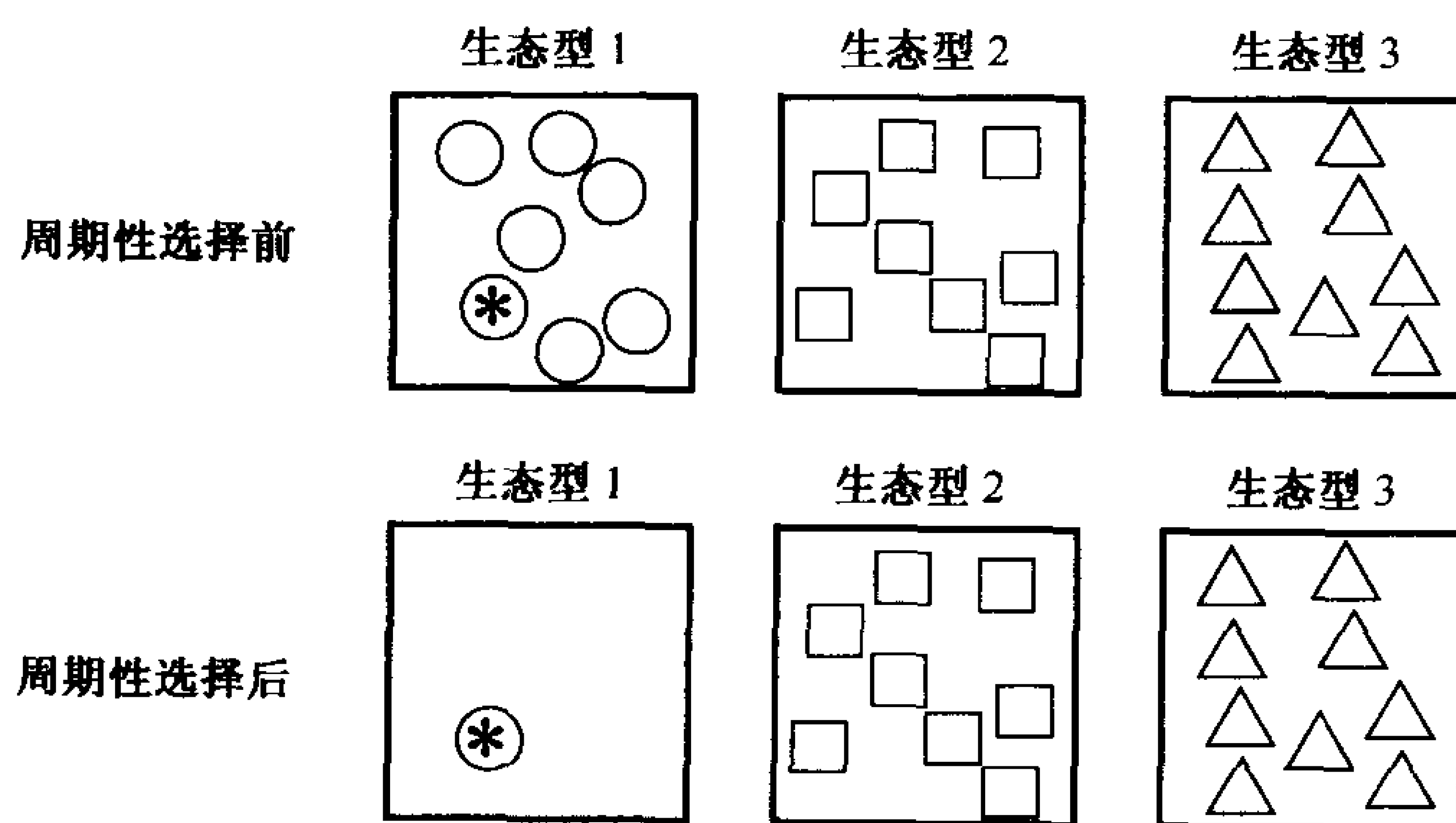


图 2 周期性选择对生态型内而非型间个体的清除作用。每个符号代表一种微生物个体，符号距离代表不同微生物序列的多样性。星号代表生态型 1 中一个适应性突变体。一旦群体差异程度加大而不受周期性选择制约（成为独立生态型），种群即发生了永久性歧变[81a]。

体，在这个群体内，周期性选择不断阻止多态性的形成，生态型间的歧变不可逆；不同生态型包括表型分化和序列分型（这点在以下关于生态型预测一节中讨论），细菌的生态型说到底是指其在生态上的差异[4,81]，根据生态型概念，细菌的“种”可以理解为在进化上由生态型特异性的周期选择所维系的世系[4]。

没有种的细菌多样性

最后，有一种观点认为，细菌分类没有动力学特性[12]，这一模型假设细菌的重组率非常高，以至周期性选择无法清除多样性，因此，在细菌种群内没有强的凝聚力，既然没有这种凝聚力，就不可能有种存在[4,12,22]。Gogarten 等[12]认为，由于高频率重组的存在，在细菌生态特异性种群中找不可靠的基因序列标记，因此，不能用序列数据对种群进行种系分析。

生态型预测和多样性中没有“种”的概念

生态型和序列分型相似性预测

没有种的概念，使得序列多样性和种群生态差异之间的相关性难以预测，但是，由于重组的存在，序列多样性只不过是微生物群体进化过程和生态特征的一个靠不住的标志[12]。相反，生态型概念可为以序列为基础细菌的多样性提供理论基础[4,5]，因为周期性选择仅在生态型内而非生态型间起作用，每个细菌的生态型最终可看成为可识别的不同与相近的其他生态型的序列簇。Palys 等[5]的研究表明，以典型细菌重组频率推算，生态型间细菌的平均序列变异要大大高于生态型内的细菌，然而，重组偶尔导致将菌株错误地划分到供体菌的生态型内，特别是只针对某一特定基因序列的分析。

Maiden 及其同事所用的 MLST 方法，在对生态型的发现和分类上，显示出比 Palys 等的方法更有效[5]，这里将重点讨论。该方法建立在周期性选择对多样性负面影响的基础上，MLST 方法中涉及的无性繁殖复合群，不易受重组的影响，因为它们一般与位

点的选择无关[3]。

我曾经设想无性繁殖复合群，即生态型，周期性选择对多样性的抑制效果导致两者的一致性[4,49,50]。因为，周期性选择的作用只限于生态型内，在两轮周期性选择之间，生态型的序列多样性只能累积到一定水平，在这段时间里，只可能发生 1 个或 2 个，至多 7 个位点的变异，不管这种变异是通过突变还是重组所产生，这就更加说明 MLST 方法中 5/7 和 6/7 标准的正确性。总之，在两轮周期性选择之间，高重组率菌株可在多位点发生变异，低重组率菌株仅通过突变累积变异，无论如何，MLST 簇与生态型对应性这一假设合理，而且可通过实验严格地验证。

相对而言，淡化了多样性中“种”的概念，就不能很好地理解周期性选择对多样性的抑制作用，并且否认任何把具体序列簇看作生态上差异种群的提法，对已定名种的多重独立序列簇存在的机制也就被否定了。

下面要考虑如何严格地证实，由周期性选择控制的 MLST 无性繁殖复合群与生态型是否有对应关系，这些种群是否无生态学意义。

生态差异预测

根据生态型概念，序列簇代表了生态差异（关于生态型内容在以下章节讨论），对这种预测的第一证据是每个序列簇与不同生境存在着对应关系。确实，一些研究表明，占据不同生态环境的一系列独立种群有极其相似序列簇[40,41,43]，例如，Ramsing 等[41]对黄石国家公园温泉中的 *Synechococus* 进行了深入研究，发现由于光照条件不同，不同深度的 *Synechococus* 有着极为相近的序列簇。

另一直接证据至少存在一些病原微生物中，其生态差异，即不同假定生态型与不同致病特性相关。MLST 方法最初是把未知病原微生物归到相应生态差异类群中，进一步研究表明，MLST 簇中有许多显示出与不同毒力和传播特性相对应[3]。

基因组学方法也能验证假定生态型与生态差异是否有对应关系，根据所含基因的差异，不同菌株可用消减杂交法（subtractive hybridization）或微阵列技术进行测试。假定生态型内菌株的生态差异，可以根据同一生态型内其他菌株是否拥有相同基因来确定，不同生态型间个体的基因差异较大[6]，例如，Salama 等[7]对幽门螺杆菌（*Holicobactor pylori*）的研究表明，可把幽门螺杆菌菌株的基因划分成多个基因簇，但这些基于基因组的基因分型，与基于 MLST 无性繁殖复合群方法确定的生态型无法形成对应关系，因为幽门螺杆菌是已知高频率重组菌株，可产生大量的无性繁殖复合群[82]。另外，利用信使核糖核酸（mRNA）微阵列技术，可以分析不同生态型间是否存在基因的差异表达[13,83]，由于基因组学技术能从生态型间不同类型中获取大量信息，其应用前景非常好[6]。

种系发生预测

一般认为，一个生态型的基因序列多样性主要受周期选择的限制，而不受遗传漂变约束，因此，任一生态型的几乎所有菌株的祖先，都可直接追溯到那些引起并在上次选择竞争中幸存的适应性突变体中[4,70]。这样一个生态种系演变史，应该与星状分化支一致，即有一个总祖先起始点，生态型各个成员间的关系彼此平行（图 3），相反，基

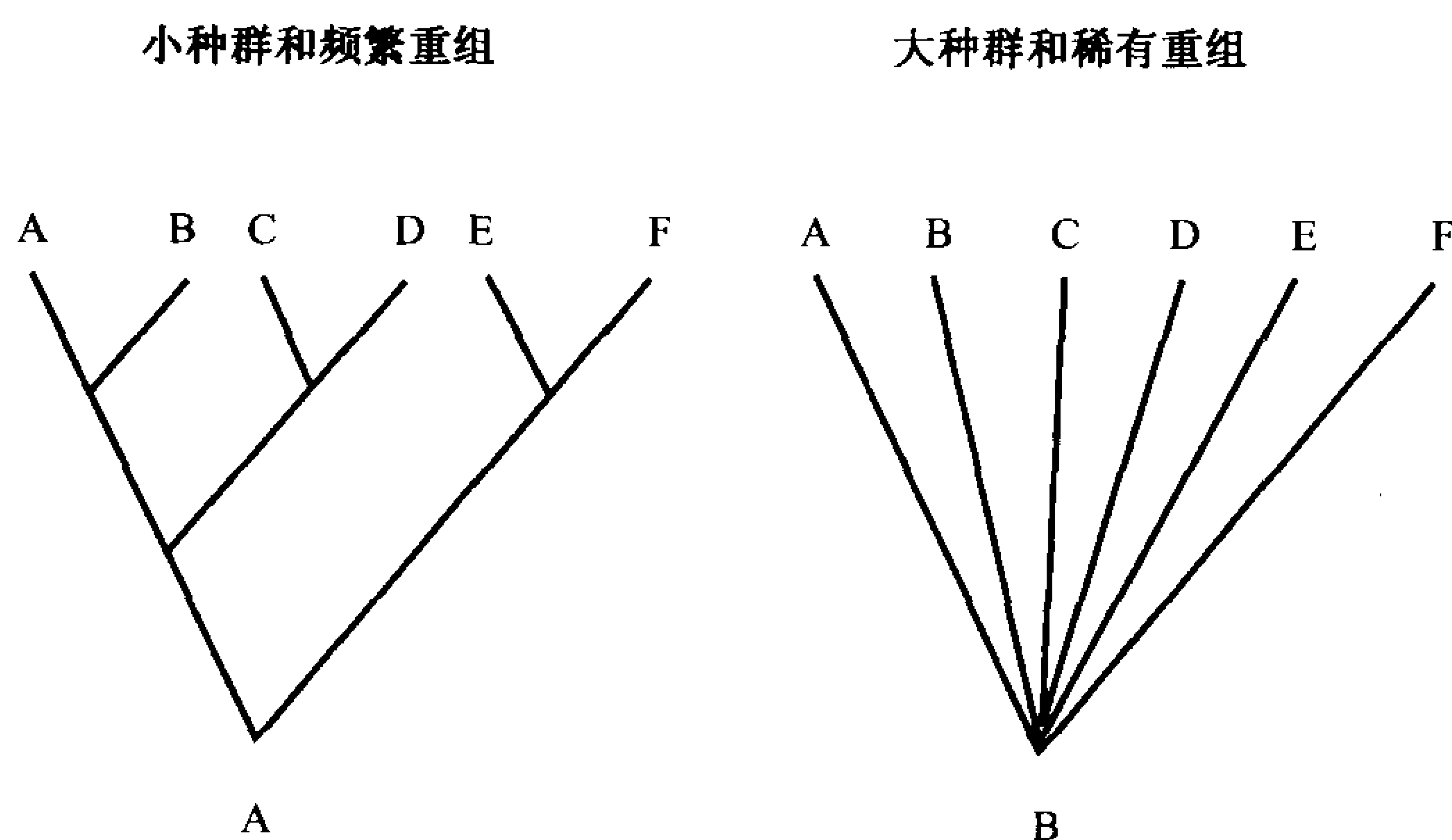


图 3　受周期性选择和遗传漂变控制多样性群体的种系发育特征[4]

因序列多样性受遗传漂变限制的种群，它的种系演变史将有许多节点。

在严格无性生态型里，基于碱基序列的种系发生，将是典型的星状分化枝，仅有由非同源相似（homoplasy）（在不同谱系中有核苷酸的趋同替代）引起的个别例外[4]。然而，重组率适中（特别与其他生态型）生态型的碱基序列种系发生关系，可能会显著偏离完美的星状分化枝。我和 Libsch 研发了一个计算机模拟系统 Star，用来测定以多基因座为基础的种系发生与星状分化枝的接近程度[4,84]。根据分类突变和重组参数，该模拟系统说明，单一生态型菌株的种系发生，可能仅有一个显著节点（即一个完美的星状），而不是两个、三个、四个或更多。

对重组率最低细菌之一的金黄色葡萄球菌（*S. aureus*）[51,52]，一个生态型应该只有一个节点[4,84]；而重组率最高细菌之一的脑膜炎奈瑟氏球菌，生态型种系发生的预测，有一个或两个明显节点，但不会有三个或更多。

我们看看 MLST 无性繁殖复合群与星状种系发生系统预测的吻合程度，我已检验了脑膜炎奈瑟氏球菌 10 种无性繁殖复合群的种系发生，是否与单一生态型的 Star 模拟系统的期望相符，结果表明，除一个无性繁殖复合群外，所有种的种系发生都只有一个或两个节点，正如根据分类重组参数所预料的那样。同样，金黄色葡萄球菌 26 种无性繁殖复合群中除一个外，其他所有种的种系发生都只含有一个明显节点，与期望值相符。两个无性繁殖复合群的混合群体中都含有两个明显节点，这表明两个无性繁殖复合群代表两个生态型。然而也有例外，有三个无性繁殖复合群混在一起，它们的种系发生只含有一个明显节点，这表明它们是同一生态型的成员。我以前曾说过，若用更高标准分析金黄色葡萄球菌（如 6/7 基因相同），得到的无性繁殖复合群可能与生态型的种系发生期望值更加相符[4]。

总之，MLST 无性繁殖复合群似乎通过了生态型检验，至少它们中的一部分在生态上显著不同，而且单一生态无性繁殖复合群的种系发生一般与它的生态型相符。

生态型自我周期性选择预测

生态型概念预测每种生态型都经历过自身周期性选择，为了验证这个预言，我们来

研究 Guttman 和 Dykhuizen[71]的碱基顺序，寻求大肠杆菌周期性选择的证据。大多数基因属四种主要序列簇，且所有菌株 *gapA* 附近染色体区域的基因序列都异常相似。他们将这种现象解释为大肠杆菌中一个适应性变体（或重组），超越其他所有变体而存活下来是不对的，就像上面讨论过的。大肠杆菌种系在生态上是多样的，因此，一个单一适应性突变体，不可能在整个物种范围内都优越，而且，即使大肠杆菌所有菌株都是一个单一生态型，周期性选择将清除整个染色体上发生的所有歧变（图 1)。

我和 Majewski[27]提出了“适应全部，作用局部（adapt globally, act locally)”模型，来解释小染色体区域附近基因的异常相似现象，就像大肠杆菌 *gapA* 那样。我们认为，大肠杆菌可能有许多生态型（也许与四种主要序列簇或更小亚簇相对应)，而 *gapA* 附近的适应性突变，一般会对所有生态型都有用，适应性突变首先引起原始生态型中歧变的清除，然后通过重组将适应性突变传递给其他生态型，促使每个生态型清除局部歧变。

“适应全部，作用局部”模型之所以这样命名，因为适应性突变（即等位基因）提高了所有生态型的适合度，但适应性突变体（即细胞）是通过在它自己生态型的内部成员中竞争生存下来的[72]。因此，对与适应性突变（基因伴随着适应性突变在生态型间传递）相关的基因，周期性选择在生态型之间或内部都会几乎彻底地清除歧变，若基因不与适应性突变相关，选择作用只会清除生态型内部发生的变异。每当小段染色体在品系间发生同化时，这个模型预测，一般有用的适应性突变从一种生态型传到另一种生态型，引起每一种生态型发生局部周期性选择。

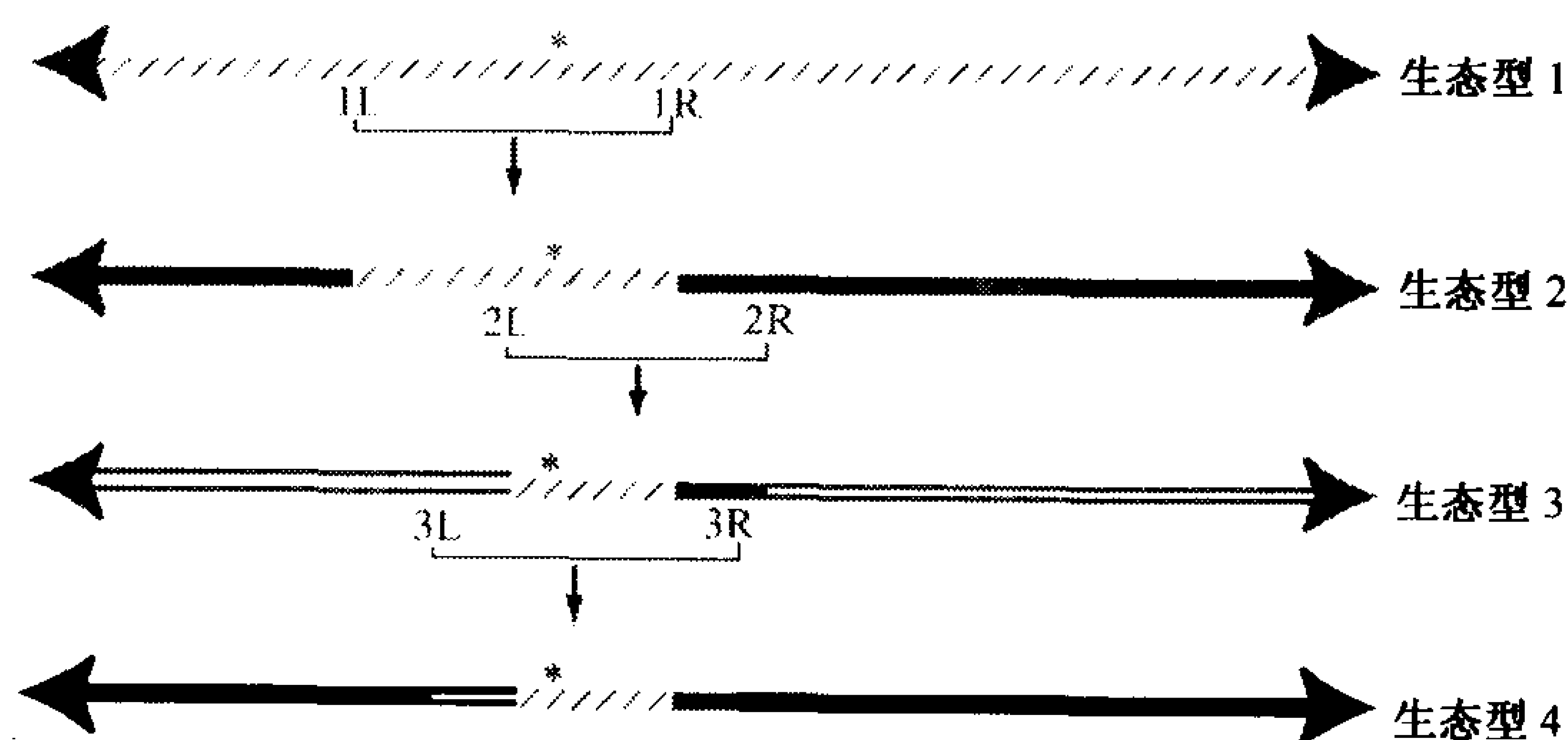


图 4　一般认为，在“适应全部，作用局部”的模型中，每对生态型间的同质化区域不同。生态 1 型首先开始发生适应性突变（标星号)，经过选择性淘汰，染色体上适应性突变附近的一小段区域（在 1L 和 1R 之间）进入生态 2 型，引起生态型选择性淘汰。然后，生态 2 型染色体上适应性突变附近一小段区域又进入生态 3 型，在那里再选择性淘汰，这样不断变化下去，最终，整个生态型适应性突变附近的序列变得相似，但每对生态型的相似（即同质化）区域界线不同（经同意引用文献[70])。

最近，我提出用基因组方法验证“适应全部，作用局部”的模型[70]。虽然，Guttman 和 Dykhuizen[71]的周期性选择模型的建立是由于偶然发现了好基因座，但今天通过整个基因组比较，容易得到他们的异常相似的染色体岛，该岛的侧翼与其他序列区域相接。“适应全部，作用局部”的模型预测，促使所有生态型发生周期性选择的适应性等位基因，是通过各自独立的基因重组而传递给每一个生态型。因此，每对生态型同

形化区域稍有不同，这反映了适应性突变在生态型之间传递的重组连接不同[70]（图4），这也许能预测，如果MLST无性繁殖复合群与生态型相对应，每对序列相似区域的连接将是独特的，这个结果进一步证实无性繁殖复合群实际上是独立的生态型，尽管由同一等位基因引起，它们都经历了各自的周期性选择。

多样性中的无种概念，对周期性选择中的同质连接没有专门预测[12]，可能是重组非常频繁，许多幸存品系在每次重组中都将获得适应性突变，这样的物种里每对菌株都有独特的同质连接点。

区分生态型和地理型

最后，在用碱基顺序方法发现和分类生态多样性时，必须注意不同序列簇并不是地域的分离所造成[5,70,85]，那些生态类型相同而分布在不同地域的种群，能分别形成各自独立的序列簇（Papke等叫作地理型，见文献[85]）。即使同一地区的细菌序列群列出，有时也很难排除地理型的假说。因为，不同地域的地理型可能近期迁入同一地区（例如通过人类运输），现在生长在同一地区的不同地理型，也许并没有经过这样一个周期选择作用，即在生态型内部对地理型多样性的纯化选择[70]。

排除地理型假设的可能途径是寻找一种适应性突变，它能促使每个假定的生态型发生各自的周期选择作用，正如上面描述的那样[70]，这也表明，单一周期选择并不能清除假定生态型的多样性，这些生态型在生态上确实截然不同。

重组能破坏生态上不同种群的分子信号吗?

生态型概念预测序列簇是独立的生态型，它在生态学上显著不同，像个星状的种系发生，并且这些生态型都经历了它们自己的周期选择作用。前两种预测在某种程度上已成功验证，并且似乎进一步巩固了生态型概念，最后一种预测还有待我介绍基因组方法的应用。

人们企图发现生态上不同种群及其紧密相关生物体的真正种系发生关系的断言是徒劳的，这时多样性无种概念是正确的[12]，Gogarten等[12]引用Feil等[67]的话：“长期而相关频繁的重组作用，消除了基因树上的种系发生信号，导致细菌种间谱系关系应该用网而不是树”，很好地说明了这个事例。然而，当推测重组也能消除生态上不同种群的碱基序列信号时，无种概念似乎又是错误的。但Gogarten并没有提到，由MLST无性繁殖复合群的划分非常可靠，而且不受所选基因和重组的影响，甚至在频繁重组的脑膜炎奈瑟氏球菌中也是这样[3]。那也是说，就重组而言，生态型和它们产生的序列群稳定，纵然是生态型中种系发生关系不够稳定，这也是为什么序列多样性能帮助我们证明生态上显著不同的种群。

基因组学的建议

幸亏有了基因组学，细菌学家正准备在细菌多样性的发现及描绘上进行深入研究，根据基因组中的基因和共享基因的表达水平，能发现生态多样性，也能证明负责入侵新生态位的基因，甚至可以追踪每个基因的供体菌株[6]。然而，在进行这项巨大的探险

活动前，要特别注意哪些基因组的差异决定了生态位的差异，因此，必须努力试图去辨别生态上相同与不同的菌株。

最后，基因组方法应该用来检测生态型的存在，就像这里解释的，分析碱基序列来发现生态型，我认为首先根据多基因座序列簇（用 MLST），把品种分为假定生态型，然后检测这些假定生态型在生态上、种系发生上和基因组上是否真独立存在。关于这方面的验证，MLST 克隆的复合体和生态型已表现出很好的一致性，进一步验证仍很重要。基因组能确证生态型概念，在科学上作出了几个重要贡献：用微排列方法显示假定生态型间基因表达模式的差异；用消减杂交和微排列来鉴定那些也许能说明假定生态型间差异本质的基因产物；通过基因组比较来证明假定生态型经历过自身的周期性选择。

一旦用碱基序列方法能成功地把生物种分为各种生态型，一个典型物种包含许多生态型，该如何继续用基因组学研究多样性呢？生态型概念应该从关注一个物种各个品系基因的多样性[73,74]，转移到关注生态型中基因组的差异，允许物种内基因组多样性的存在。应该知道，假定生态型的基因含量和基因表达的变化，仅仅能代表一个种群在两次周期选择间的随机变化，但是，导致不同生态型形成的基因组差异，负责入侵新生态位和维持种群的共存，这些才是重要的值得我们注意的基因组差异。

致谢

很感激 Michael Dehn 帮助理清原稿的许多建议。本项研究受国家科学基金会DEB－9815576 的资金和 Wesleyan 大学研究基金的资助。

（周世力，朱晨光　译）

参考文献

1. DeLong, E, Pace N. Environmental diversity of bacteria and archaea. Syst Biol 2001; 50:470–478.
2. Ward DM, Weller R, Bateson MM. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 1990; 345:63–65.
3. Maiden MC, Bygraves JA, Feil E, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 1998; 953140–3145.
4. Cohan FM. What are bacterial species? Annu Rev Microbiol 2002; 56:457–487.
5. Palys T, Nakamura LK, Cohan FM. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int J Syst Bacteriol 1997; 47:1145–1156.
6. Joyce EA, Chan K, Salama NR, Falkow S. Redefining bacterial populations: a post-genomic reformation. Nat Rev Genet 2002; 3:462–473.
7. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc Natl Acad Sci USA 2000; 97:14668–14673.
8. Hihara Y, Kamei A, Kanehisa M, Kaplan A, Ikeuchi M. DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. Plant Cell 2001; 13:793–806.
9. Harrington CA, Rosenow C, Retief J. Monitoring gene expression using DNA microarrays. Curr Opin Microbiol 2000; 3:285–291.

10. Nesbo CL, Nelson KE, Doolittle WF. Suppressive subtractive hybridization detects extensive genomic diversity in *Thermotoga maritima.* J Bacteriol 2002; 184:4475–4488.
11. Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res 2001; 11: 1404–1409.
12. Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 2002; 19:2226–2238.
13. Feldgarden M, Byrd N, Cohan FM. Gradual evolution in bacteria: evidence from *Bacillus* systematics. Int J Syst Bacteriol 2003; 149:3565–3573.
14. Sneath PH. Future of numerical taxonomy. In: Goodfellow M, Jones D, Priest F (eds). Computer-Assisted Bacterial Systematics. Orlando, FL: Academic, 1985, pp. 415–431.
15. Rossello-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev 2001; 25:39–67.
16. Goodfellow M, Manfio GP, Chun J. Towards a practical species concept for cultivable bacteria. In: Claridge MF, Dawah HA, Wilson MR (eds). Species: The Units of Biodiversity. London: Chapman and Hall, 1997.
17. Johnson J. Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. Int J Syst Bacteriol 1973; 23:308–315.
18. Wayne LG, Brenner DJ, Colwell RR, et al. Report of the Ad Hoc Committee on reconciliation of Approaches to Bacterial Systematics. Int J Syst Bacteriol 1987; 37:463–464.
19. Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA:DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 1994; 44:846–849.
20. de Queiroz K. The general lineage concept of species, species criteria, and the process of speciation. In: Howard DJ, Berlocher SH (eds). Endless Forms: Species and Speciation. Oxford, UK: Oxford University Press, 1998, pp. 57–75.
21. Meglitsch P. On the nature of species. Syst Zool 1954; 3:491–503.
22. Templeton A. The meaning of species and speciation: a genetic perspective. In: Otte D, Endler J (eds). Speciation and Its Consequences. Sunderland, MA: Sinauer, 1989, pp. 3–27.
23. Mayr E. Animal Species and Evolution. Cambridge, MA: Belknap Press of Harvard University Press, 1963.
24. Mayr E. Systematics and the Origin of Species from the Viewpoint of a Zoologist. New York: Columbia University Press, 1944.
25. Simpson G. Principles of Animal Taxonomy. New York: Columbia University Press, 1961.
26. Wiley E. The evolutionary species concept reconsidered. Syst Zool 1978; 27:17–26.
27. Eldredge N. Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Thought. New York: Oxford University Press, 1985.
28. Yohalem DS, Lorbeer JW. Intraspecific metabolic diversity among strains of *Burkholderia cepacia* isolated from decayed onions, soils, and the clinical environment. Antonie Van Leeuwenhoek 1994; 65:111–131.
29. Logan NA, Berkeley RC. Identification of *Bacillus* strains using the API system. J Gen Microbiol 1984; 130(Pt 7):1871–1882.
30. Lan R, Reeves PR. Gene transfer is a major factor in bacterial evolution. Mol Biol Evol 1996; 13:47–55.
31. Edwards RA, Olsen GJ, Maloy SR. Comparative genomics of closely related Salmonellae. Trends Microbiol 2002; 10:94–99.
32. Parkhill J, Achtman M, James KD, et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 2000; 404:502–506.
33. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escheri-*

chia coli O157:H7. Nature 2001; 409:529–533.
34. Read TD, Brunham RC, Shen C, et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 2000; 28:1397–1406.
35. Tettelin H, Saunders NJ, Heidelberg J, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
36. Alm RA, Ling LS, Moir DT, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 1999; 397:176–180.
37. Feil EJ, Maiden MC, Achtman M, Spratt BG. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol Biol Evol 1999; 16: 1496–1502.
38. Feil EJ, Smith JM, Enright MC, Spratt BG. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics 2000; 154:1439–1450.
39. Dykhuizen DE, Green L. Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 1991; 173:7257–7268.
40. Rocap G, Distel DL, Waterbury JB, Chisholm SW. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 2002; 68:1180–1191.
41. Ramsing NB, Ferris MJ, Ward DM. Highly ordered vertical structure of *Synechococcus* populations within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. Appl Environ Microbiol 2000; 66:1038–1049.
42. Ward DM. A natural species concept for prokaryotes. Curr Opin Microbiol 1998; 1:271–277.
43. Beja O, Koonin E, Aravind L, et al. Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. Appl Environ Microbiol 2002; 68:335–345.
44. Vulic M, Lenski RE, Radman M. Mutation, recombination, and incipient speciation of bacteria in the laboratory. Proc Natl Acad Sci USA 1999; 96:7348–7351.
45. Lawrence JG. Gene transfer in bacteria: speciation without species? Theor Popul Biol 2002; 61: 449–460.
46. Cohan FM. The effects of rare but promiscuous genetic exchange on evolutionary divergence in prokaryotes. Am Naturalist 1994; 143:965–986.
47. Posada D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol Biol Evol 2002; 19:708–717.
48. Smith JM, Smith N, O'Rourke M, Spratt BG. How clonal are bacteria? Proc Natl Acad Sci USA 1993; 15:4384–4388.
49. Cohan FM. Clonal structure: an overview. In: Pagel M (ed). Encyclopedia of Evolution. Vol. 1. New York: Oxford University Press, 2002, pp. 159–161.
50. Cohan FM. Population structure and clonality of bacteria. In: Pagel M (ed). Encyclopedia of Evolution. Vol. 1. New York: Oxford University Press, 2002, pp. 161–163.
51. Feil EJ, Cooper JE, Grundman H, et al. How clonal is *Staphylococcus aureus?* J Bacteriol 2003; 185:3307–3316.
52. Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, Spratt BG. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). Proc Natl Acad Sci USA 2002; 99:7687–7692.
53. Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J Bacteriol 2000; 182: 1016–1023.
54. Majewski J, Cohan F. M. DNA sequence similarity requirements for interspecific recombination in Bacillus. Genetics 1999; 153:1525–1533.
55. Vulic M, Dionisio F, Taddei F, Radman M. Molecular keys to speciation: DNA polymorphism

and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci USA 1997; 94:9763–9767.

56. Duncan KE, Ferguson N, Kimura K, Zhou X, Istock CA. Fine-scale genetic and phenotypic structures in natural populations of *Bacillus subtilis* and *Bacillus licheniformis*: important implications for bacterial evolution and speciation. Evolution 1994; 48:2002–2025.
57. Majewski J. Sexual isolation in bacteria. FEMS Microbiol Lett 2001; 199:161–169.
58. Cohan FM. Sexual isolation and speciation in bacteria. Genetica 2002; 116:359–370.
59. Majewski J, Cohan FM. The effect of mismatch repair and heteroduplex formation on sexual isolation in Bacillus. Genetics 1998; 148:13–18.
60. Rao BJ, Chiu SK, Bazemore LR, Reddy G, Radding CM. How specific is the first recognition step of homologous recombination? Trends Biochem Sci 1995; 20:109–113.
61. Rayssiguier C, Thaler DS, Radman M. Thc barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature 1989; 342:396–401.
62. Doolittle WF. Lateral genomics. Trends Cell Biol 1999; 9:M5–M8.
63. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299–304.
64. Licht TR, Christensen BB, Krogfelt KA, Molin S. Plasmid transfer in the animal intestine and other dynamic bacterial populations: the role of community structure and environment. Microbiology 1999; 145(Pt 9):2615–2622.
65. Maynard Smith JM, Dowson CG, Spratt BG. Localized sex in bacteria. Nature 1991; 349:29–31.
66. Arthur W. Mechanisms of Morphological Evolution: A Combined Genetic, Developmental and Ecological Approach. New York: Wiley, 1984, pp. 182–186.
67. Feil EJ, Holmes EC, Bessen DE, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci USA 2001; 98:182–187.
68. Atwood KC, Schneider LK, Ryan FJ. Periodic selection in *Escherichia coli*. Proc Natl Acad Sci USA 1951; 37:146–155.
69. Cohan FM. Genetic exchange and evolutionary divergence in prokaryotes. Trends Ecol Evol 1994; 9:175–180.
70. Cohan FM. Periodic selection and ecological diversity in bacteria. In: Nurminsky D (ed). Selective Sweep. Georgetown, TX: Landes Bioscience, in press.
71. Guttman DS, Dykhuizen DE. Detecting selective sweeps in naturally occurring *Escherichia coli*. Genetics 1994; 138:993–1003.
72. Majewski J, Cohan FM. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics 1999; 152:1459–1474.
73. Lan R, Reeves PR. Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 2000; 8:396–401.
74. Boucher Y, Nesbo CL, Doolittle WF. Microbial genomes: dealing with diversity. Curr Opin Microbiol 2001; 4:285–289.
75. Ravin A. The origin of bacterial species: genetic recombination and factors limiting it between bacterial populations. Bacteriol. Rev 1960; 24:201–220.
76. Ravin A. Experimental approaches to the study of bacterial phylogeny. Am Nat 1963; 97:307–318.
77. Linz B, Schenker M, Zhu P, Achtman M. Frequent interspecific genetic exchange between commensal Neisseriae and *Neisseria meningitidis*. Mol Microbiol 2000; 36:1049–1058.
78. Cohan FM. Does recombination constrain neutral divergence among bacterial taxa? Evolution 1995; 49:164–175.
79. Giraud A, Matic I, Tenaillon O, et al. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. Science 2001; 291:2606–2608.

80. Denamur E, Lecointre G, Darlu P, et al. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell 2000; 103:711–721.
81. Cohan FM. Bacterial species and speciation. Syst Biol 2001; 50:513–524.
81a. Cohan FM. The role of genetic exchange in bacterial evolution. ASM News 1996; 62:631–636.
82. Suerbaum S, Smith JM, Bapumia K, et al. Free recombination within *Helicobacter pylori*. Proc Natl Acad Sci USA 1998; 95:12619–12624.
83. Gibson G. Microarrays in ecology and evolution: a preview. Mol Ecol 2002; 11:17–24.
84. Cohan FM, Libsch J. Sequence-based evidence for numerous ecologically distinct and irreversibly separate taxa within bacterial species.
85. Papke T, Ramsing NB, Bateson MM, Ward DM. Geographical isolation in hot spring cyanobacteria. Environ Microbiol 2003; 5:650–659.

12

病原菌和共生菌与寄主的协同进化

Robert A. Feldman

引言

许多微生物与后生动物紧密生活在一起，它们之间的关系可以是专性的，如细胞器和人类细胞内许多病原菌；也可以是共栖的和近乎边缘性的，对微生物或寄主既没有正面影响也没有负面影响。通常，微生物与寄主之间有长期共进化历史，协同进化的结果是微生物与寄主间紧密的生理和调节依赖性。基因组学通过辨别微生物病原性基因及由寄主生理活动所调节的遗传调控系统，在解码微生物与寄主关系的遗传学基础方面起关键作用。不久的将来，基因组学技术将使我们了解到基因调控和微生物与寄主间代谢耦联的类型，这些研究又会促进对微生物与寄主协同进化历史的进一步了解。

本章将阐述一些微生物与寄主协同进化的机制，包括微生物与寄主协同进化的起源、持续及其协同进化的维持。并描述不同类型微生物与寄主之间的关系。从微生物与细胞器开始，再到微生物与植物、昆虫共生体、海洋无脊椎动物和人类之间的相互作用，最后，讨论未能培养的人体内微生物区系（microflora）和基因组学，为未来这方面的研究带来希望。

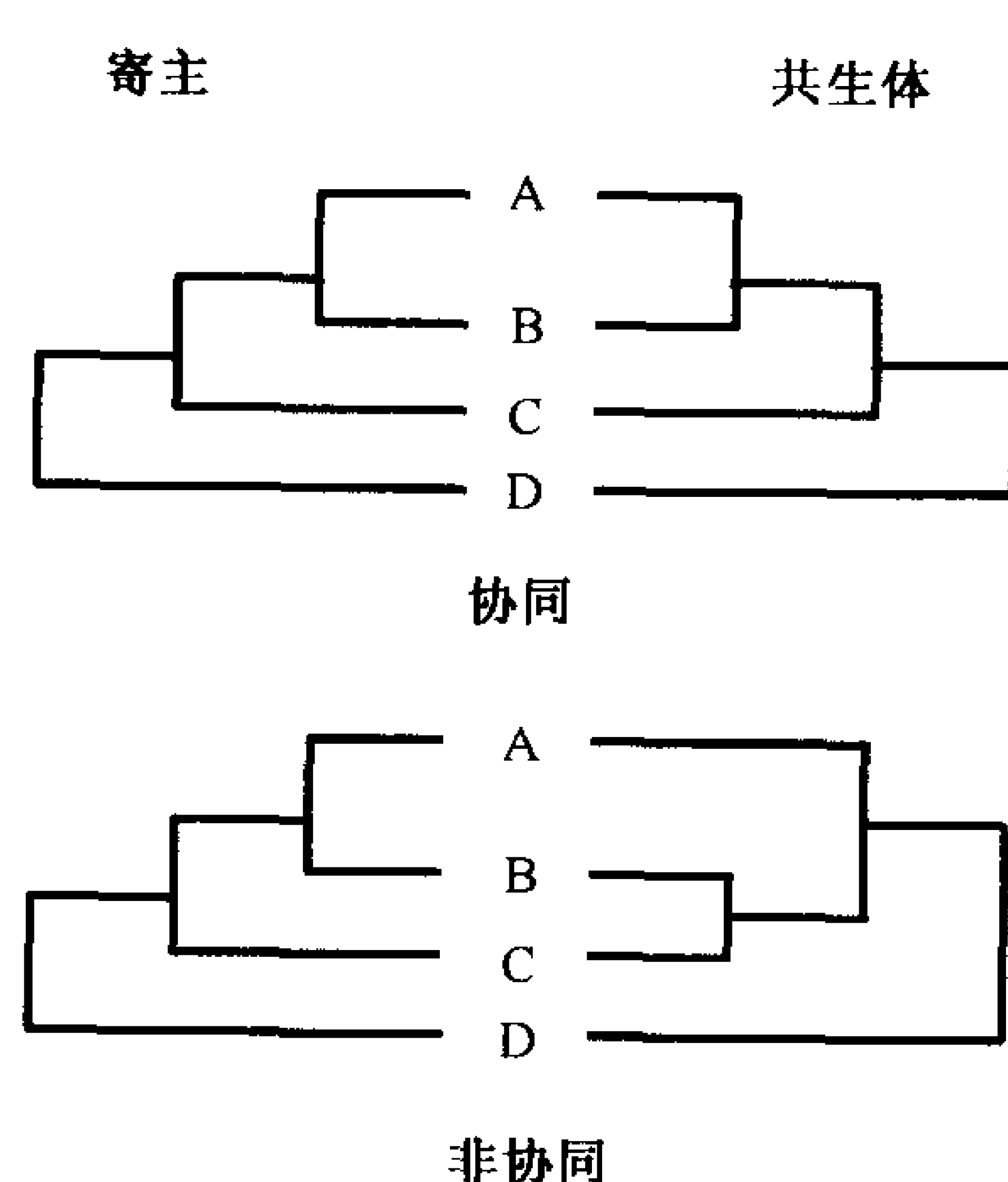


图 1　协同与非协同的种系发生。图示的基因树由假设的寄主及其共生体的基因计算获得。在上图中，寄主及其共生体通过分享相同的进化史产生协同性；在下图中，寄主及其共生体种系发生的不同，显示它们经历了不同的进化史。

协同进化机制：持久、恢复和选择

共生体的产生和传播策略

共生体的建立和通过寄主的传播有三种机制，分别是横向机制（从已建立起共生关系的成体传递到新成年个体）、纵向机制（由亲代传递给子代）和复感染机制（环境引起的），每种共生发生传播都可由种系发生的协同性或共生体与寄主间基因进化史的非协同性来反映（图 1）。

在横向传播机制中，共生体的种系发生可由独立于寄主种系进化史之外的亲缘地理信号来解释。内共生体的横向传播发生在深海的衣管虫中，这可由寄主管虫及其共生体的非协同性种系发生树来证明[1]。

在纵向传播机制中，传播是通过配子从亲代传给子代，纵向传播可由寄主－共生体的种系发生的协同性（共同的进化史）反映出来，纵向传播使深

海蛤类和化能自养细菌之间形成物种共同进化的关系（协同种系发生）[2]。

共生体传播中的复感染模式机制要难解释得多，但是共生体的种系发生树应该表现出明显的亲缘地理因素。

细胞通讯

很多不同种类的植物和动物病原菌，都采用Ⅲ型分泌系统将效应蛋白传递进寄主细胞以辅助它们侵入寄主细胞。Ⅲ型分泌系统由20种蛋白质组成，它们与鞭毛的组成蛋白有亲缘关系[3]，Ⅲ型分泌系统的机制目前还不清楚，但已知道，该过程涉及一个针状复合体，正是它作为中空的导管让Ⅲ型分泌蛋白通过[3]。一旦Ⅲ型分泌蛋白进入寄主细胞的细胞质中，这些蛋白质就能够调节寄主细胞的功能。通过纵向机制传递的象鼻虫，其互利内共生体中就存在有Ⅲ型分泌系统，据推测，此内共生体至少起源于5000万年前[4]。

群体感应（quorum sensing）是指细菌根据种群密度对基因表达做出的调控[5]，通过自诱导过程起作用。细菌产生一种可扩散的化合物（一种自诱导因子），当种群抗扩增时，这种化合物便会在环境中积累，细菌通过这种化合物来感知和监控环境中的种群密度[5]。控制群体感应的基因表达系统，在自由生活的海洋哈氏弧菌（*Vibrio harveyi*）中发现，该菌具有双组分感应系统，每个系统都由专一感知因子和自诱导因子组成，二者都对密度依赖的生物荧光基因表达起控制作用[6]。哈氏弧菌Ⅰ型自诱导因子（AI-1，一种高丝氨酸形成的内酯）及其产生过程中所涉及的基因（*luxL* 和 *luxM*）已了解得很清楚[5~7]。在哈氏弧菌、大肠杆菌和鼠伤寒沙门氏菌中，系统2的自诱导受最近才发现的称作 *luxS* 高度同源基因所控制[6]。对这些群体感应系统之间相似性和差异性的愈加全面了解，可提供用抗生素控制病原菌的新方法。

基因组简并和突变率

微生物共生体和病原菌生活在一个静态、富含代谢物、并由寄主不断提供营养物的环境中，病原菌的许多基因变得无用而被丢失了。总之，相对于自由生活的个体，共生或病原的生活方式经常会导致基因的丢失和整体基因组的萎缩。许多丢失基因的功能被寄主补偿，进一步基因组丢失会导致代谢灵活度下降，此时，微生物限制在不可逆的专性寄生生活方式中。到达极限时，严格的细胞内寄生生物有着已知最小的基因组[8]，DNA修复基因的丢失和碱基组成偏嗜 A + T 都会加大突变率。

共生体和病原菌所组成的种群过小会导致进化速率的上升，由于环境的限制，重复瓶颈效应和极小有效种群，对自由生活的微生物，突变体会更容易在共生体中固定下来[9,10]。共生菌和病原菌的氨基酸替换率，可以高达自由生活物种的两倍[11]，病原菌的高突变率产生了高水平基因组序列多样性，这种多样性可用来建立病原菌近期遗传起源，还可用来研究病原菌传播的流行病学[12]，这不仅与生物武器威胁有特别关系，而且还与传染病爆发有关。

微生物-寄主关系的类型

细胞器：线粒体和叶绿体

通常会接收这么一种观点，现代质体和线粒体同真核细胞的关系，源于寄主细胞对曾自由生活细菌的接纳。在所有已知的共生现象中，质体（光合作用的细胞器）和线粒体（产能的细胞中发生呼吸作用的细胞器）是最古老和高度发达的共生体[13]。这些细胞器高度简并的基因组，必需基因的转移和对核基因的调控，是互利关系最强有力的证据[14]。很难想像如果没有这些细菌与其寄主的早期“联姻”，出现在地球上的生命会进化成什么样子。有关质体和叶绿体起源的关键性问题，不再是它们是否被寄主摄入，而是分多少次摄入？原始寄主又是什么生物？

对于原始质体，越来越多分子水平的种系发生证据支持这一观点：蓝细菌和真核寄主细胞间的单种系，起源变化产生了现代绿色植物、红藻和灰藻[14,15]。Moreira 等[15]对核内延长因子 2 和 RNA 聚合酶Ⅱ氨基酸序列的种系发生研究显示，红藻和绿色植物同属于真核生物一个单种系发生的进化枝。其他分子水平种系发生的研究结果也支持这一观点：一是对核糖体大小亚单位 rRNA 序列的研究[16]；另一是对四个核基因（α-微管蛋白、β-微管蛋白、肌动蛋白、延伸因子 1-α）的综合研究[17]。质体次级共生式摄入（摄入到一个已经拥有了质体的寄主细胞中）的次数仍不清楚[14]。

rRNA 种系发生表明，真核细胞线粒体起源于一种类似于 α-多形杆菌的祖先[18]，从现今进化角度看与线粒体的祖先关系最密切的物种后裔是由立克次体分支的一组细胞内寄生生物，包括以下几个属：立克次体、无形体属和埃立克次体[19]。现在强有力的证据证明，线粒体有单种系起源，这些证据包括由 rRNA 基因和蛋白质序列推导的进化树一致，并且，所有线粒体基因都是一个已知最大线粒体基因组的一部分[19]。

线粒体基因组大小和基因数高度多变，最小线粒体基因组（小于 6kb）在疟原虫中发现，最大线粒体基因组(366 924bp)在开花植物拟南芥中发现[19]。具鞭毛原生动物 *Reclinomonas americana* 的线粒体基因组，在目前所有测序的线粒体中拥有最多的基因数（97 个基因)[20]。*R. americana* 的线粒体基因组除含有其他线粒体基因组的所有蛋白质编码基因外，还含有 18 个新基因、一个 5S rRNA 基因、一个 RNAse P 基因和一个细菌型 RNA 聚合酶基因。因此，*R. americana* 的线粒体基因组代表着与细菌线粒体最类似的基因组[19,20]。

由于内共生生活方式，细胞器经历了基因组的大量丢失，并把功能基因转移给寄主细胞核。这些基因多为管家基因，它们参与诸如氨基酸生物合成、核苷生物合成、无氧酵解和调控等功能[19]。现在看来，从动物线粒体向核内的基因转移已经停止，这是因为核基因组和线粒体基因组是不同遗传密码。在显花植物中，线粒体基因组和核基因组是完全一样的遗传密码。从植物线粒体向核内的基因转移十分活跃[21,22]。从叶绿体向核内的基因转移比较常见，还可直接在实验室中观察到[23]。

在杂色藻（*Chromophyte algae*）中，细胞内分区、转运和基因产物的共享似乎达到进化顶峰。杂色藻起源于对红藻或绿藻（已有叶绿体和细胞核）的次级共生，由具鞭毛但不能进行光合作用的寄主提供细胞核、内膜和线粒体[24,25]。在进化为隐藻类的过

程中，作为内共生体红藻的细胞核不断萎缩，变成了残体，被称为类核体（nucleomorph）[25]。现今的隐藻通过四个基因组（叶绿体、细胞核、类核体和线粒体），协同完成基因产物在复杂内膜系统中的转移[25]。

植物病原体和共生体

在多形杆菌门α-多形杆菌纲中，有些细菌有与真核细胞发生紧密生理关系的倾向。这个纲里的根瘤菌（和豆科植物形成的一种能固氮的共生体）、土壤杆菌（植物病原菌）和立克次体（动物细胞内病原菌）是进化树上很集中的一簇[26]。

根癌土壤杆菌是一种植物病原菌，它是植物冠瘿病的成因，能将一段特定自身基因组 DNA 片段，整合到真核寄主细胞的基因组中去。在自然界中，这种现象发生在根癌土壤杆菌对植物伤口处细胞释放的小分子信号物的感知[27]，在这些分子的诱导下，根癌土壤杆菌的 Ti 质粒 *vir* 基因编码的产物激活了一系列基因的表达，导致 T-DNA 转移到植物细胞中，T-DNA 随机整合到寄主植物细胞的基因组中。T-DNA 基因的表达可以改变植物生长激素的水平，导致植物冠瘿瘤的形成，T-DNA 基因的表达还刺激植物产生一种 opines 物质供细菌生长。两个研究组分别独立地对根癌土壤杆菌进行了测序，并几乎同时报道[27,28]。根癌土壤杆菌基因组的大小为 5.67Mb，由四部分组成：一条长 2.841Mb 环状染色体、一条长 2.075Mb 线状染色体和两种质粒——大小为 542.8kb 的 pAtC58 和大小为 214.2kb 的 pTiC58[27,28]。线状染色体的端粒可通过发夹环结构以共价键形式封闭[28]，参与植物细胞转化和根瘤形成的基因分散在所有 4 个基因组部件中。在根癌土壤杆菌和已测序的所有根瘤菌的基因组中，发现了大量 ABC 转运基因，这说明在充满竞争的土壤环境中，需要有高亲和力的转运系统来保证养分的获取[28]。

苜蓿根瘤菌（*Sinorhizobium meliloti*）是一种与苜蓿形成固氮共生体的α-多形杆菌。作为根瘤菌成员，它感染植物根部形成根瘤并能在根瘤中固氮（即将氮元素转化成可被植物利用的形式）[29]。在根瘤形成过程中，根瘤菌和植物体保持通讯，还建立起代谢共享网络，以便使根瘤菌能从植物体获取含碳化合物，同时为植物提供固定的氮[30]。苜蓿根瘤菌基因组由 3.65Mb 染色体、1.35Mb 大质粒 pSymA 和 1.68Mb 大质粒 pSymB 组成[30]，总共 6.7Mb 的基因组有 6204 个编码蛋白的可读框[30]。氮代谢基因在大质粒 pSymA 上成簇排列，而大质粒 pSymB 含有参与小分子转运的基因。将苜蓿根瘤菌基因组和百脉根瘤菌（*Mesorhizobium loti*）基因组进行比较[31]，发现百脉根根瘤菌中仅 35%基因与苜蓿根瘤菌直系同源，而苜蓿根瘤菌的三重基因组携带的遗传信息分散在百脉根根瘤菌中[30]，这有力证明，根瘤菌在基因数和基因组结构上差异显著。然而，根癌土壤杆菌和苜蓿根瘤菌的环状染色体有高度同线性，这支持了环状染色体起源于最初类似于α-多形杆菌祖先的观点[27,30]。

昆虫微生物共生体

昆虫与各种各样的细菌共生，其中很多细菌已研究得比较透彻[34]。许多共生菌在种系发生上都被划分在γ-多形杆菌的肠杆菌科中，并与大肠杆菌有较近的亲缘关系。这些共生菌存在于胸喙亚目（半翅目）昆虫中，包括木虱、蚜虫、粉蚧和粉虱[33]，这些系统中研究得最清楚的是蚜虫与黑草属（*Buchnera*）细菌形成高度互利的胞内共生

体。黑草属细菌亲缘关系最近的祖先是大肠杆菌，而黑草属细菌基因组似乎主要通过基因组精减从一株与大肠杆菌类似的祖先进化而来[34,35]。

Buchnera aphidicola 基因组的主要特点是基因组较小（640 681bp），含有为寄主合成必需氨基酸的基因，但缺乏合成寄主非必需氨基酸的基因[35]。黑草属细菌缺乏编码细胞表面的受体基因、调控基因和防御基因[35,36]，这就意味着它的基因组完全就是一个共生的、胞内寄居型的基因组。两株来源于不同寄主 *Aphidicola pisum* 和麦二叉蚜（*Schizaphis graminum*）的 *B.aphidicola* 菌株基因组极其稳定。从 5000 万～7000 万年前二者在进化上产生分支以后，它们的基因组中都没有任何重排现象（包括倒位、易位、重复、基因获得）[37]，二者的基因组大小（641.5kb，640.7kb）、编码蛋白基因的数目（545，564）以及鸟嘌呤和胞嘧啶的含量（G+C）（26.2，26.3）都很相似[37]。二者基因组的明显相似性表明，*Buchnera* 基因组精减早在趋异进化前就在二者共同祖先中发生了[34]。然而，用脉冲电泳技术对 5 个亚科蚜虫的 9 个 *Buchnera* 菌株的基因组大小进行测定，结果表明 *Buchnera* 基因组的大小从 450～670kb 不等[38]。显然，*Buchnera* 的比较基因组学仍需进一步研究。

海洋无脊椎动物与微生物之间的共生

海绵的古生菌共生体

在加州海岸第一次发现 Crenarchaeal 共生体，它在常见的海洋生物 *Axinella mexicana* 体内生活[39]，该共生体与嗜高温的 Crenarchaeal 在种系发生上有亲缘关系，但却在中温海域出现，且能在 10℃ 水族馆中的海绵体内生长。此温度要比实验室任何一株 Crenarchaeal 的培养温度低 60 多度[39]。这种共生体的发现，使人们开始认识到适冷性和适温性 Crenarchaeal 在环境中的普遍存在。此共生体命名为 *Crenarchaeum symbiosum*[39]，还不能在海绵组织外培养。自发现 *C.symbiosum* 后，越来越多的古生菌被发现与海绵[40]和其他无脊椎动物共生在一起[41]。

Stein 等发明的种系发生定位技术评估了 Crenarchaeum 群落基因组的多样性[42]，该技术涉及建立大（最初大约 40kb，现在超过 150kb）[43]插入克隆的环境 DNA 文库，再用探针来筛选目的基因（通常是 rRNA 基因），然后对含 rRNA 目的基因的克隆进行全长测序，并以 rRNA 基因为基准，用计算机程序搜索两侧的可读框，把它与数据库中的条目进行比较（BLAST 是最流行的方法）[44]，就可得出菌株在种系发生上的推论。这种钓鱼式的途径最成功时，可为未能培养微生物提供代谢途径信息和生理信息[45]。

通过对重叠 40kb 的 fosmid 克隆对 *C.symbiosum* 进行测序。Schleper 等对海绵体内微生物群落的三个 fosmid 克隆进行了克隆、测序和组装[46]。在组装的 fosmid 克隆中，序列对比很整齐的一些区域显示出高度序列多样性，其他区域却显示出基因组重排的迹象。有趣的是，整个群落的 16S rRNA 序列多样性还不到 3%，这项研究还揭示这个古生菌共生体群落在启动子区域上的差异[46]，首次对未能培养近缘共生菌的大片段基因组的插入序列进行了比较，使其对近缘微生物基因组的多样性有了初步认识。

深海热泉微生物共生体

在东太平洋海丘火山热泉处，栖息的主要一类特化群体属 *Alvinellidae* 科多毛纲

(*polychaetes*)环节动物 *Alvinellid*，可产生复杂的几丁质管状结构供生活在“黑烟囱”周围的几百种动物居住。烟囱喷涌出高达450℃超热水，而这些动物就生活在距烟囱仅几厘米的地方。*Alvinellid* 虫长 2.5cm，表面被细丝覆盖，生活在后生动物温度最高的环境中，它们居住的管体可达 80℃，有时甚至达 105℃[47]。它们与 40～50 种 ε 多形杆菌生活在一起，这些蠕虫与微生物共存，在蠕虫细丝覆盖的背上包了一层细菌，尚不清楚细菌对寄主有什么特殊优势。推测可能与耐热性、对毒性重金属的抵抗力和利用细菌作为一种直接食物有关。

正在开展对 *Alvinellid* 虫细菌群落的泛基因组研究，将大量测序环境克隆，通过与数据库比对，对它们进行初步鉴别。不同基因和相同基因的变异体排布于微阵列玻片上，用来检测其他的环境信息文库。这将成为功能环境基因组的一部分，希望能据此给群体中的基因作生理和代谢功能的指定。

深海管虫化能自养共生体

在东太平洋海丘 2500 米深的热泉处，出现了另一个专化的深海微生物-无脊椎动物共生现象。这种身长 2.5 米庞大的管虫（*Riftia pachyptila*），是已知最长的海生无脊椎动物，管虫的成虫无嘴、内脏和肛门，完全依赖内共生细菌提供营养[48]，管虫有高度发达的循环和神经系统，并利用专一性氧硫可逆结合血红蛋白运输氧、二氧化碳和 HS 供给共生菌。共生菌居住在称为“营养体（trophosome）”的特化器官中。在营养体中，共生菌利用卡尔文循环进行化能自氧反应，把二氧化碳合成苹果酸，并将四碳糖和五碳糖提供给寄主组织。

共生菌是 γ 多形杆菌纲的成员，与大肠杆菌有亲缘关系，寄主通过横向转移从环境中摄入细菌，因此这些细菌有自由生活阶段。共生体几乎没有亲缘地理学特性，仅表现出低水平的遗传多样性，在种系发生方面与寄主无协同性[1,49]。共生菌生活史中有自由生活阶段，并有鞭毛蛋白基因[50]和组氨酸激酶信号系统[51]。目前尚不清楚幼年管虫和共生菌之间的识别机制是什么，但这是正在进行基因组研究计划的焦点之一。

人类病原体

流感嗜血菌（*Haemophilus influenzae*）是第一个获得基因组全序列的生物，它的完成展示了基因组时代的到来。除了技术上的巨大成就外，其基因组提供了生物学几个方面的知识，其一是流感嗜血菌非病原株 Rd 与病原株 b 型株在感染性方面有差异[52]，特别是在 Rd 株中完全缺乏，编码 8 个纤毛基因的黏附基因簇（使细菌黏附到寄主细胞上），因此，可以像这样通过全基因组分析来推断和提出病原生物学的新假说。

1983 年发现幽门螺旋杆菌（*Helicobacter pylori*）是引起人类胃溃疡的病原微生物[53]，它生活在低 pH（2.0～3.0）的胃内环境中，它利用胃中分子氢作为能源[54]。世界人口几乎一半受到感染，是人类最常见的病原体。幽门螺旋杆菌基因组揭示了作为极端环境中的病原体的几个有趣特征，它有高度发达的限制性修饰系统，存在 5 种以上不同机制可以吸附到寄主胃表皮细胞及阳性膜内电压。

幽门螺旋杆菌有几种增加抗原变化的方法，含有大量二核苷酸重复，允许滑链错配

增加基因型和表型变化[55]。它的表面脂多糖分子模仿人血细胞抗原，使它比其他肠道细菌有较小的免疫原性[55]。包括幽门螺旋杆菌基因的种系树经常有很“奇怪”（即与16S rRNA不同）的布局，这表明该菌在多形杆菌的进化中较早地产生了分支，经历了快速进化的过程，有高水平的水平基因转移，常以拟纵向转移的方式从母细胞传给子细胞。因此，一些幽门螺旋杆菌基因适合作为研究人类迁移的标记。

支原体（*Mycoplasmas*）是低G+C含量、完全缺乏细胞壁的革兰氏阳性菌，是严格人类病原体。对人类而言，肺炎支原体（*Mycoplasmas pneumoniae*）引起非典型性肺炎。生殖道支原体（*Mycoplasma genitalium*）引起非淋球性尿道炎。这两个有亲缘关系支原体的基因组首批完成了测序，因此，可以将它们进行比较[58,59]。两个基因组都很小（肺炎支原体为816kb、生殖道支原体为580kb），因此基因含量也很少。生殖道支原体所含基因最少（517），是目前已知基因组最小的独立存在的生物体。支原体有高度简化的代谢基因系统，通过从寄主中转运营养以满足需求。

苍白密螺旋体（*Treponema pallidum*）是引起人类性病梅毒的螺旋菌。螺旋菌和人类有长期复杂的共同进化历史，自从欧洲人到达美洲以后，它就多次导致原住民的大量死亡。螺旋菌是仅存于人体内的严格病原体，和其他螺旋菌，如引起莱姆关节炎的布氏疏螺旋体（*Borrelia burgdorferi*）相比，苍白密螺旋体的基因组较小(1 138 006bp)，G+C含量为52.8%，编码1041个ORF[61]，含有与布氏疏螺旋体和生殖道支原体似的全套DNA复制基因。同布氏疏螺旋体一样，苍白密螺旋体含有多种运输蛋白，但缺少呼吸电子传递链基因。这两个基因组都含有保守的运动和趋化基因，苍白密螺旋体的外膜几乎无膜蛋白，可能是逃避寄主识别。

衣原体（*Chlamydiae*）是引起多种疾病的细胞内病原体，一篇文章比较了两种衣原体的基因组（分别是感染鼠的沙眼衣原体（*Chlamydia trachomatis*）MoPn和感染人的肺炎衣原体（*Chlamydia pneumoniae*）AR39）[62]，两基因组都很小（沙眼衣原体为1 042 519bp，编码894个ORF，肺炎衣原体为1 230 230bp，编码1052个ORF），有高度同线形，推测是由于严格细胞内病原体的生态环境所造成[62]，仅几个基因决定寄主种类的定向，或许是那些参与核苷摄取和代谢的基因。比较基因组学还揭示肺炎衣原体存在噬菌体，该噬菌体可能在发病机制方面发挥作用。

脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）是一种β-多形杆菌，引起脑膜炎和败血症，其基因组全长2 272 351bp，编码2158个ORF[63]。它通过相变异控制基因表达，躲避寄主的免疫系统，基因组揭示了49个（除以前知道的16个以外）可能发生相变异的基因，包括参与外膜蛋白、磷脂糖、菌毛、限制性修饰、荚膜形成和摄取铁离子的基因[63]。

空肠弯杆菌（*Campylobacter jejuni*）是存在于食物中的病原体，种系发生学归为Δ-多形杆菌。它是世界上细菌性食物中毒的主要原因，尤其是含家禽和饮水的细菌性中毒[64]。基因组全长1 641 481bp，编码1654个蛋白质，这使它成为密度最高的细菌基因组[65]。在编码表面结构生物合成和修饰的基因中，有一些由短重复序列组成的高度可变部分[65]。尽管空肠弯曲杆菌和幽门螺旋杆菌在种系发生和生理上有很近的亲缘关系，但它们的基因组几乎没有共同之处（主要是管家基因的功能），这可能是强大选择压（驱使它们生存在不同环境中）造成的[65]。

很多低G+C含量的革兰氏阳性链球菌的全基因组已经测出，包括化脓链球菌（*Streptococcus pyogenes*）和肺炎链球菌（*Streptococcus penumoniae*）[66,67]。化脓链球菌是专性人类病原体，引起一系列疾病，包括喉咙痛、猩红热、脓疱病、丹毒、蜂窝织炎、败血病、中毒性休克综合征、食肉病、风湿热和急性血管球性肾炎[66]。肺炎链球菌每年引起全球一百万以上的人死亡，是美国排列前十名的死亡原因之一[67]。两种基因组都有很多毒力相关基因，很多编码细胞表面蛋白或分泌蛋白。它们的基因组还显示与生活在同一环境中的革兰氏阴性细菌间有高水平的水平基因转移。化脓链球菌有4种不同编码超级抗原蛋白的原噬菌体。

弧菌是广泛存在于海洋和淡水系统中自由生活的γ多形杆菌，它们作为鱼类[68]和头足类动物[69]的共生体存在。在发展中国家和地区，霍乱弧菌（*Vibrio cholerae*）是人类重要的病原体，霍乱流行时，严重的痢疾可引起大量人口死亡。病原学上，环境和气候因素可能在霍乱弧菌从无害水生菌转变为恶性病原菌的过程中起重要作用[70]。霍乱弧菌基因组有两个2.96Mb和1.07MB环形染色体，共编码3885个ORF[71]，较大染色体上（染色体1）的大多数基因对正常细胞功能和致病性是必需的，而染色体2包括大部分假定基因和一个存在质粒上的整合子岛基因捕获系统[71]。种系发生学上，染色体1可能和大肠杆菌相关，染色体2则可能起源于细胞对一个大质粒的捕获[71]。

绿脓杆菌（*Pseudomonas aeruginosa*）是普遍存在于沿海海洋系中的γ多形杆菌。它在岩石和土壤上形成生物膜，可侵染植物、动物组织，是对抗生素有高度抗性的人类机会病原菌。绿脓杆菌在烧伤病人、导尿管病人、医院获得性感染性肺炎病人中，引起几乎无法治疗的顽固性感染，并且是囊肿性纤维病人的主要死亡原因[72]。绿脓杆菌基因组较大（6.3Mb），编码5570个ORF，包括很多并系同源基因家族[72]，其中之一是编码外膜蛋白的150个基因，这些外膜蛋白是不同环境中抗生素抗性竞争的进化结果。

金黄色葡萄球菌（*Staphylococcus aureus*）是低G+C含量的革兰氏阳性菌（与杆菌、梭菌、支原体平行的分类单元），是医院获得性感染的主要原因，能引起中毒性休克综合征和葡萄球菌猩红热，也能导致皮肤感染、食物中毒、肺炎、脓血症、骨髓炎和感染性心内膜炎[73]。过去40年金黄色葡萄球菌对抗生素的抗性不断增强，现在一些菌株对所有抗生素有抗性。基因组测序计划对两个金黄色葡萄球菌抗性菌株进行了比较，一个抗新青霉素（菌株N315）和一个抗万古霉素（菌株Mu50），它们的基因组大小为2 813 641bp和2 878 040bp，总G+C百分比为32.8和32.9，ORF数量为2595和2697，全序列相似性（96%）很接近[73]。抗生素抗性基因存在由噬菌体编码的抗性岛内，其基因组也有致病基因岛（pathogenicity island，PAI），PAI是编码很多致病因子的大段病原基因组DNA（10～200kb）（见第4章），这些因子包括黏附素、毒素、侵袭素、蛋白分泌系统、铁离子摄入系统和其他[74]。PAI也许能积累基因并能通过水平转移从一个分类单元转移到另一个分类单元。金黄色葡萄球菌还包含几个编码高度变化的外毒素超级表面抗原蛋白的转座子、插入序列、所有已知的和70个新致病因子[73]。

人类微生物组，非培养的共生体：另一个人类基因组

人类基因组草图序列的完成，像人类首次登陆月球一样是科学史上的里程碑。它的

完成应看成是基因组时代的开始，而不像人们所说的那样已经进入了后基因组时代。人类基因组提供了一套正常基因的框架，一张草图，它能解释人类进化和生物学的起始点，而在人类进化和生物学中一个重要的、并到现在还尚被忽略的部分是对人体微生物区系组成的理解。

人体内是一个复杂的环境，有500～1000种微生物共同生活，它们中至少有一半还不能人工培养。这些微生物总称为微生物组（microbiome），应作为人类基因组的一部分包括在内[75]。假设这些微生物中的每个基因组为5Mb，而每个基因组编码4000个基因，那将意味着人体内有大约2～4百万个非人类基因，这表明微生物的非人类基因可能是人类基因的100多倍[75]。这些微生物的基因产物有的可分泌出来，对人类细胞生理学产生影响。此外，微生物产生代谢产物，有时产生毒素，这就出现问题了：在正常人类生理学中，这些微生物组基因及其基因产物的作用是什么?

研究16S rRNA基因多样性的技术手段，使微生物生态学经历了一场革命。这些研究试图对微生物群体的所有成员进行测序，然后将数据与已获培养微生物的数据进行比较，从而确定出未能培养微生物种系的从属关系。这些微生物分子生态学研究，已应用于人类未能培养微生物区系的研究中，并取得了大开眼界的结果。用这种方法对人齿龈下沟微生物区系进行研究，发现了50多种独特种类，分属5个大类的细菌[76]，这些未能培养微生物与疾病的特定关系尚不清楚。最成功的研究是将幽门螺旋杆菌和胃溃疡联系在一起[77]，但未能培养微生物还可能与很多疾病有关，包括心血管疾病、前列腺炎、克罗恩氏病、肾结石、脑神经疾病妥瑞症、糖尿病、多发性硬化、热带性腹泻、川崎病、韦格内肉芽肿、类肉状瘤病、风湿性关节炎、乳腺癌、非何杰金氏淋巴瘤、睾丸癌和前列腺癌等。

系统研究人体内未能培养微生物区系，可以了解个体之间的生理差异，提供研究环境影响的工具，并帮助阐明大量非遗传病因的疾病。到目前为止，很多研究都围绕rRNA的序列特征[76,78]，而新技术，如微阵列和transcript-filtering subtraction可以更快地收集数据，从而给出新药物靶位点、预防性措施和个体化药物。

结论

随着从培养和未能培养生物体中积累更多的基因组数据，生物体基因组和生理学之间的关系会更清晰。总之，基因组有独特的生态和进化史，基因组有残存、镶嵌和合成的特点，现已开始从群落的角度认识基因组学，而不在乎基因产物来自何种生物，此外，也开始把基因组学看作是泛基因组与环境在生理意义上的相互作用。目前把基因组学比作蛋白质研究中的X射线衍射结晶技术，该技术可提供许多蛋白质结构中的某一种，同样地，基因组学将使我们能开始理解生物与群落的生理学、生态学和进化学。

（徐进平　译）

参考文献

1. Feldman RA, Black MB, Cary CS, Lutz RA, Vrijenhoek RC. Molecular phylogenetics of bacterial endosymbionts and their vestimentiferan hosts. Mol Marine Biol Biotechnol 1997; 6:268–277.
2. Peek AS, Feldman RA, Lutz RA, Vrijenhoek RC. Cospeciation of chemoautotrophic bacteria and deep sea clams. Proc Natl Acad Sci USA 1998; 95:9962–9966.
3. Galan JE, Collmer A. Type III secretion machines: bacterial devices for protein delivery into host cells. Science 1999; 284:1322–1328.
4. Dale C, Plague GR, Wang B, Ochman H, Moran NA. Type III secretion systems and the evolution of mutualistic endosymbiosis. Proc Natl Acad Sci USA 2002; 99:12,397–12,402.
5. Fuqua WC, Winans SC, Greenberg EP. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. J Bacteriol 1994; 176:269–275.
6. Surette MG, Miller MB, Bassler BL. Quorum sensing in *Escherichia coli*, *Salmonella typhimurium*, and *Vibrio harveyi*: a new family of genes responsible for autoinducer production. Proc Natl Acad Sci USA 1999; 96:1639–1644.
7. Bainton NJ, Bycroft BW, Chhabra SR, et al. A general role for the lux autoinducer in bacterial cell signalling: control of antibiotic biosynthesis in *Erwinia*. Gene 1992; 116:87–91.
8. Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995; 270:397–403.
9. Funk DJ, Wernegreen JJ, Moran NA. Intraspecific variation in symbiont genomes: bottlenecks and the aphid-buchnera association. Genetics 2001; 157:477–489.
10. Maruyama T, Kimura M. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc Natl Acad Sci USA 1980; 77:6710–6714.
11. Itoh T, Martin W, Nei M. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc Natl Acad Sci USA 2002; 99:12,944–12,948.
12. Read TD, Salzberg SL, Pop M, et al. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 2002; 296:2028–2033.
13. Margulis L. Symbiosis in Cell Evolution. San Francisco: Freeman, 1981.
14. Palmer JD. A single birth of all plastids? Nature 2000; 405:32–33.
15. Moreira D, Le Guyader H, Philippe H. The origin of red algae and the evolution of chloroplasts. Nature 2000; 405:69–72.
16. Van der Auwera G, Hofmann CJ, De Rijk P, De Wachter R. The origin of red algae and cryptomonad nucleomorphs: a comparative phylogeny based on small and large subunit rRNA sequences of *Palmaria palmata*, *Gracilaria verrucosa*, and the *Guillardia theta* nucleomorph. Mol Phylogenet Evol 1998; 10:333–342.
17. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 2000; 290:972–977.
18. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. Mitochondrial origins. Proc Natl Acad Sci USA 1985; 82:4443–4447.
19. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science 1999; 283:1476–1481.
20. Lang BF, Burger G, O'Kelly CJ, et al. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 1997; 387:493–497.
21. Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 2000; 408:354–357.
22. Gray MW. Mitochondrial genes on the move. Nature 2000; 408:302–305.
23. Huang CY, Ayliffe MA, Timmis JN. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 2003; 422:72–76.
24. Cavalier-Smith T. Membrane heredity and early chloroplast evolution. Trends Plant Sci 2000;

25. Douglas S, Zauner S, Fraunholz M, et al. The highly reduced genome of an enslaved algal nucleus. Nature 2001; 410:1091–1096.
26. Woese CR. Bacterial evolution. Microbiol Rev 1987; 51:221–271.
27. Goodner B, Hinkle G, Gattung S, et al. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. Science 2001; 294:2323–2328.
28. Wood DW, Setubal JC, Kaul R, et al. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science 2001; 294:2317–2323.
29. Downie JA, Young JPW. The abc of symbiosis. Nature 2001; 412:597–598.
30. Galibert F, Finan TM, Long SR, et al. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 2001; 293:668–672.
31. Kaneko T, Nakamura Y, Sato S, et al. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res 2000; 7:331–338.
32. Moran NA. Bacterial menageries inside insects. Proc Natl Acad Sci USA 2001; 98:1338–1340.
33. Clark MA, Baumann L, Thao ML, Moran NA, Baumann P. Degenerative minimalism in the genome of a psyllid endosymbiont. J Bacteriol 2001; 183:1853–1861.
34. Moran NA, Mira A. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol 2001; 2:2–54.
35. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. Nature 2000; 407:81–86.
36. Moran NA, Wernegreen JJ. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol Evol 2000; 15:321–326.
37. Tamas I, Klasson L, Canback B, et al. Fifty million years of genomics stasis in endosymbiotic bacteria. Science 2002; 296:2376–2379.
38. Gil R, Sabater-Munoz B, Latorre A, Silva FJ, Moya A. Extreme genome reduction in *Buchnera* spp: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci USA 2002; 99: 4454–4458.
39. Preston CM, Wu KY, Molinski TF, DeLong EF. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen nov sp nov. Proc Natl Acad Sci USA 1996; 93:6241–6246.
40. Webster NS, Watts JEM, Hill RT. Detection and phylogenetic analysis of novel crenarchaeote and euryarchaeote 16S ribosomal RNA gene sequences from a Great Barrier Reef sponge. Marine Biotechnol 2001; 3:600–608.
41. van Hoek AH, van Alen TA, Sprakel VS, et al. Multiple acquisition of methanogenic archaeal symbionts by anaerobic ciliates. Mol Biol Evol 2000; 17:251–258.
42. Stein JL, Marsh TJ, Wu KY, Shizuya H, Delong EF. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bacteriol 1996; 178:591–599.
43. Beja O, Suzuki MT, Koonin EV, et al. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ Microbiol 2000; 2:516–529.
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol 1990; 215:403–410.
45. Beja O, Aravind L, Koonin EV, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 2000; 289:1902–1906.
46. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 1998; 180:5003–5009.
47. Chevaldonne P, Desbruyeres D, Childress JJ. ... and some even hotter. Nature 1992; 359:593.
48. Cavanaugh CM, Gardiner SL, Jones ML, Jannasch HW, Waterbury JB. Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila* Jones: possible chemoautotrophic symbionts. Science 1981; 213:340–341.
49. Di Meo CA, Wilbur AE, Holben WE, Feldman RA, Vrijenhoek RC, Cary SC. Genetic variation

among endosymbionts of widely distributed vestimentiferan tubeworms. Appl Environ Microbiol 2000; 66:651–658.

50. Millikan DS, Felbeck H, Stein JL. Identification and characterization of a flagellin gene from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. Appl Environ Microbiol 1999; 65:3129–3133.
51. Hughes DS, Felbeck H, Stein JL. A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. Appl Environ Microbiol 1997; 63:3494–3498.
52. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
53. Warren JR, Marshall B. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet 1983; 1:1273–1275.
54. Olson JW, Maier RJ. Molecular hydrogen as an energy source for *Helicobacter pylori*. Science 2002; 298:1788–1790.
55. Tomb JF, White O, Kerlavage AR, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997; 388:539–547.
56. Falush D, Wirth T, Linz B, et al. Traces of human migrations in *Helicobacter pylori* populations. Science 2003; 299:1582–1585.
57. Spratt BG. Microbiology: Stomachs out of Africa. Science 2003; 299:1528–1529.
58. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 1996; 24: 4420–4449.
59. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res 1997; 25:701–712.
60. Hutchison CA, Peterson SN, Gill SR, et al. Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 1999; 286:2165–2169.
61. Fraser CM, Norris SJ, Weinstock GM, et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 1998; 281:375–388.
62. Read TD, Brunham RC, Shen C, et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 2000; 28:1397–1406.
63. Tettelin H, Saunders NJ, Heidelberg J, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
64. Blaser MJ. Epidemiologic and clinical features of *Campylobacter jejuni* infections. J Infect Dis 1997; 176:S103–S105.
65. Parkhill J, Wren BW, Mungall K, et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 2000; 403:665–668.
66. Ferretti JJ, McShan WM, Ajdic D, et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 2001; 98:4658–4663.
67. Hoskins J, Alborn WE Jr, Arnold J, et al. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 2001; 183:5709–5717.
68. Haygood MG, Distel DL. Bioluminescent symbionts of flashlight fishes and deep-sea anglerfishes form unique lineages related to the genus *Vibrio*. Nature 1993; 363:154–156.
69. Nyholm SV, Stabb EV, Ruby EG, McFall-Ngai MJ. Establishment of an animal-bacterial association: recruiting symbiotic vibrios from the environment. Proc Natl Acad Sci USA 2000; 97: 10,231–10,235.
70. Colwell RR. Global climate and infectious disease: the cholera paradigm. Science 1996; 274: 2025–2031.
71. Heidelberg JF, Eisen JA, Nelson WC, et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 2000; 406:477–483.
72. Stover CK, Pham XQ, Erwin AL, et al. Complete genome sequence of *Pseudomonas aeruginosa*

PA01, an opportunistic pathogen. Nature 2000; 406:959–964.

73. Kuroda M, Ohta T, Uchiyama I, et al. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 2001; 357:1225–1240.
74. Hacker J, Kaper JB. Pathogenicity islands an the evolution of microbes. Annu Rev Microbiol 2000; 54:641–679.
75. Hooper LV, Gordon JI. Commensal host-bacterial relationships in the gut. Science 2001; 292: 1115–1118.
76. Kroes I, Lepp PW, Relman DA. Bacterial diversity within the human subgingival crevice. Proc Natl Acad Sci USA 1999; 96:14,547–14,552.
77. Relman DA. The human body as a microbial observatory. Nat Genet 2002; 30:131–133.
78. Tanner MA, Shoskes D, Shahed A, Pace NR. Prevalence of corynebacterial 16S rRNA sequences in patients with bacterial and "nonbacterial" prostatitis. J Clin Microbiol 1999; 37:1863–1870.
79. Cummings CA, Relman DA. Using DNA microarrays to study host-microbe interactions. Emerg Infect Dis 2000; 6:513–525.
80. Diehn M, Relman DA. Comparing functional genomic datasets: lessons from DNA microarray analyses of host-pathogen interactions. Curr Opin Microbiol 2001; 4:95–101.
81. Weber G, Shendure J, Tanenbaum DM, Church GM, Meyerson M. Identification of foreign gene sequences by transcript filtering against the human genome. Nat Genet 2002; 30:141–142.

第五部分：微生物基因组的调查

13

植物病原菌基因组调查

C. Robin Buell

引言

植物容易被许多病原菌感染，包括病毒、细菌、真菌和线虫，植物又有天生的防御机制，并起动“从头开始（*de novo*）”的防御反应，来确保继续生存。这些防御机制对病原菌是一个挑战，而这些病原菌反过来会用致毒或致病机制，逃避寄主的防御反应，以此确保生长繁殖的营养来源。这种寄主的防御与病原菌毒力的微妙平衡，导致了病原菌和寄主植物的共同进化。

植物疾病防御的中心理论是“基因对基因（gene-for-gene）”假说。该假说最初在20世纪40年代，由Flor对亚麻锈病病原菌（*Melampsora lini*）疾病防御的遗传研究后提出的[1]。根据这个模型，寄主的单个主导基因产物与病原菌的单个无毒（avirulence）基因产物的相互作用，决定了寄主与病原菌相互作用的结果。随着分子技术的发展，克隆了多个病原菌无毒基因和寄主抗病基因，从分子水平证实了这个遗传模型（抗病研究大事记见文献［2］)。“基因对基因”模型的研究开发，要么用抗病基因，要么用无毒基因，来构建有抗病特性（包括抗细菌、真菌和病毒）的转基因植物，这意味着通过生物工程作物的新控制机制是合理的。从而通过对植物病原菌致病机制和毒力机制的研究，获得新认识并开发出控制植物病原菌的新药剂。因此，目前所有植物病原菌基因组计划的首要目的，是针对致毒和致病机制。

本文的目的是已经完成或正在进行的细菌、真菌和线虫的基因组计划，表1列出了所有细菌性植物病原菌，包括已经完成和正在进行的基因组序列计划，表2和表3分别列出了真菌、卵菌和线虫的表达序列标签（expressed sequence tag，EST）计划。可利用网页（http://www.tigr.org/～vinita/PPwebpage.html 和 http://www.oardc.ohio-state.edu/Phytophthora/genome_links.htm）了解当前植物病原菌以及与植物有关微生物基因组计划的进展状况。

表1　植物细菌性病原菌基因组计划

细菌名	基因组大小/Mb	疾病	参考文献	网址
已完成基因组				
根癌土壤杆菌 c58 (*Agrobacterium tumefaciens*)	5.7	对许多品种的冠瘿	22, 23	www.genome.washington.edu
青枯病菌 GMI1000 (*Ralstonia solanacearum*)	5.8	番茄细菌性枯萎	19	www.cereon.com

续表

细菌名	基因组大小/Mb	疾病	参考文献	网址
地毯黄单胞菌柑橘致病变种 306 (*Xanthomonas axonopodis* pv *citri* 306)	5.2	柑橘溃疡	30	http://genoma4.iq.usp.br/xanthomonas/
野油菜黄单胞杆菌野油菜致病变种 ATCC3391 (*Xanthomonas campestris* pv *campestris*)	5.1	十字花科植物黑斑	8	http://genoma4.iq.usp.br/xanthomonas/
苛养木杆菌 9a5c(*Xylella fastidiosa*)	2.7	柑橘花斑缺绿症		www.watson.fapesp.br
草图基因组				
丁香假单胞菌丁香致病变种 B728a (*Pseudomons syringae* pv *syringae*)	~6.0	豆类细菌性褐色斑		www.jgi.doe.gov
苛养木杆菌杏变种(*Xylella fastidiosa*)(Dixon)	2.6	杏叶枯萎		www.jgi.doe.gov
苛养木杆菌夹竹桃变种 (*Xylella fastidiosa*)(ann1)	2.6	夹竹桃叶枯萎		www.jgi.doe.gov
正在进行中的基因组				
密执安棒形杆菌(*Clavibacter michiganensis* spp. *sepedonicus*)	~2.5	马铃薯软烂		http://www.sanger.ac.uk/projects/microbes/
胡罗卜软腐欧氏杆菌黑腐致病变种 (*Eruwinnia carotovora* spp. *atroseptica*)	~5.0	马铃薯软烂、黑根		http://www.sanger.ac.uk/projects/microbes/
菊欧文氏菌 3937(*Eruinia chrysanthem*)	3.7	许多宿主品种软烂		www.genome.wisc.edu http://www.tigr.org/tdb/mdb/mdbinprogress.html
丁香假单胞菌蕃茄致病变种 DC3000 (*Pseudomons syringae* pv *tomato*)	6.5	番茄细菌斑		http://www.tigr.org/tdb/mdb/mdbinprogress.html
野油菜黄单胞杆菌野油菜致病变种 8004 (Xanthomonas campestris pv campestris)	5.1	十字花科植物黑斑		http://www.chgb.org.cn/en-ke.html
苛养木杆菌皮尔斯病致病株 (Xylella fastidiosa-Pierce's disease strain)	2.6	葡萄皮尔斯病		http://www.lbi.ic.unicamp.br/world/xf.grape/

表 2　植物真菌性和卵菌性病原菌表达序列标签计划

菌种名	表达序列标签数	疾病
布氏白粉禾谷类白粉菌(*Blumeria graminis* f sp *hordei*)	4908	大麦白粉病
百慕达草草皮黄叶菌(*Fusanrium sporotrichioides*)	7625	玉米谷穗霉烂
稻瘟霉菌(*Magnaparthe grisea*)	14 160	稻瘟病
小麦壳针孢枯病菌(*Mycosphaerella graminicola*)	1158	小麦壳针孢霉病叶斑病原菌
马铃薯晚疫病菌(*Phytophthora infestans*)	2129	土豆和番茄晚期枯萎
大豆疫霉(*Phytophthora sojae*)	2004	大豆根茎霉烂
共计	31 985	

注:表达序列标签数取自 2002 年 5 月 10 日公布的数据库,只包括 1000 以上表达序列标签的菌种。稻瘟霉菌(*M. grisea*)数目以 *M. grisea* 的名称贮存的表达序列标签,不包括该菌的前名 *Pyricularia grisea*。

表 3　植物线虫病原表达序列标签计划

线虫名	表达序列标签数	疾病
马铃薯白线虫(*Globodera pallida*)	1832	马铃薯孢囊线虫病
马铃薯金线虫(*Globodera rostochiensis*)	5934	马铃薯孢囊线虫病
孢囊线虫(*Heterodera glycines*)	4327	马铃薯孢囊线虫病
花生根结线虫(*Meloidogyne arenaria*)	3334	根结线虫病
北方根结线虫(*Meloidogyne hapla*)	6157	北方根结线虫病
南方根结线虫(*Meloidogyne incognita*)	10 899	南方根结线虫病
爪哇根结线虫(*Meloidogyne javanica*)	5600	根结线虫病
共计	38 083	

注：表达序列标签数取自 2002 年 5 月 10 日公布的数据库，只包括 1000 以上的表达序列标签的菌种。植物致病性线虫表达序列标签的当前数据在网上可查 http://www.nematode.net。

植物病原菌基因组

在日常生活中尽管农业很重要，但是将基因组用到植物病原菌在很大程度上仍受限制。截至 2002 年 5 月，已超过 75 个细菌基因组全序列被报道，却只有 6 个基因组属植物病原性细菌(http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html)。这 6 个基因组全序列针对 5 种植物病原菌。已完成其他几种植物细菌性病原菌的序列草图，还有其他一些细菌基因组正在测序中。目前，已开始了对植物细菌性病原菌基因组的广泛调查，更多的基因组计划将会产生一个植物病原菌的多样性数据。

苛养木杆菌

最先完成的植物细菌性病原菌基因组是苛养木杆菌（*Xylella fastidiosa*）基因组，由巴西“核酸测序和分析组织（Organization for Nucleotide Sequencing and Analysis, ONSA)”完成，于 2000 年 7 月公布[8]。苛养木杆菌是限制木质部的细菌，可导致几种重大经济损失的疾病[9,10]，特别是菌株 9a5c，它是柑橘花斑缺绿症的病原[11]，可导致感染柑橘植物萎缩和感染叶坏死[12]。细菌由昆虫（如 sharpshooter）传播，当前的控制方法是修剪感染植株，通过使用杀虫剂减少或消除昆虫媒介[12]。除了控制苛养木杆菌的媒介传播以外，它是一种很难在实验室操作的病原菌，因此，很难在生理和物种水平上了解该病原菌和疾病。苛养木杆菌基因组的完成是植物病理研究的主要转折点，也是通过基因组促进对植物病理深入了解的成功例证。

苛养木杆菌菌株 9a5c 的基因组包括 2.68Mb 染色体和两个大小分别为 51.1kb、1.28kb 的质粒，它们的 G+C 百分含量分别是 52.7%、49.6% 和 55.6%[8]。该染色体和两个质粒共编码 2848 个可读框（ORF)。根据数据库中已知基因的相似性，可以推测出共有 1314（46%）个可读框的功能，其代谢和转运体系可从基因组注释中推测，并与本质部的限制性营养环境一致[8]。

从对 9a5b 基因组序列分析中，对该病原菌的生理和毒理有了进一步认识。在一些革兰氏阴性病原菌中，名为超敏感反应和致病调控元基因（*hypersensitive response* and *pathogenicity*, *hrp*）对细菌在植物中的生长是必须的，*hrp* 基因编码分泌系统类型Ⅲ组分，通过横跨内外膜的分泌器官，分子被传到寄主细胞，从而决定植物与微生物相互作

用结果（对分泌系统类型Ⅲ的近期综述见文献［13～15］)。通过类型Ⅲ分泌传输的一群分子被认为是效应子，并包括来自无毒基因的产物。令人吃惊的是，9a5c 基因组中并没有无毒基因或分泌系统类型Ⅲ基因，这说明苛养木杆菌的寄主特异性并不是由典型的“基因对基因”模型决定[8]。然而，在 9a5c 基因组中，已确定了估计对毒理和病理有作用的其他一些基因，包括纤维素酶基因和涉及将细菌黏附到寄主细胞壁的蛋白质基因，例如，胞外多糖产物，类型四菌毛纤维及非菌毛黏附[8]，这些发现为该品种的毒理机制进行系统研究提供了平台。

苛养木杆菌皮尔斯菌株是许多疾病的病原，包括柑橘花斑缺绿症、葡萄皮尔斯病、假桃病，以及包括杏树、李树、橡树和枫树在内的几种寄主的叶枯萎[9,10]。通过典型的基因型鉴定技术，如限制性长度片段多态性研究和随机扩增 DNA，对该品种的多样性进行了鉴定[16]。为了进一步阐明该品种的多样性和寄主特异性，能源部联合基因组研究所（Department of Energy Joint Genome Institute, DOE-JGI），对苛养木杆菌的另外两个来自夹竹桃和杏的附加菌株进行了序列分析，见数据库：

http://www.jgj.doe.gov/gi.microbial/html/xylella-oleadey/xyle-olccn-homepage. html 和 http://www.jgj.doe.gov/JGI-microbial/html/xylella-almond/xyle-almnd-homepage.html。

虽然这些基因组的测序还没有最终完成，但是，根据另外两个寄主品种的两个菌株序列，可以理解该菌种的多样性，并通过比较基因组研究，提供了潜在侵染特定寄主品种的病理或毒理必需的序列。除这两个序列草图外，巴西 ONSA 研究组已分析了能引起葡萄皮尔斯病的苛养木杆菌一个菌株(http://www.ibi.ic.unicamp.br/world/xf-grape/)，因此，总共有苛养木杆菌 4 个菌株的全基因组序列可以利用。

青枯病菌

另一个有意义的革兰氏阴性植物细菌性病原菌是能引起维管枯萎的病原菌青枯病菌(*Ralstonia solanacearum*)。该细菌能感染维管组织的木质部管道，由于细菌大量繁殖和胞外多糖产物的积累，造成管道堵塞，植物不能通过管道组织输送水分而导致萎缩[17]。

选择分析的是菌株 GMI1000，它能感染双子叶植物蔬菜番茄，还能感染模式植物拟南芥（*Arabidopsis thaliana*)[18]。与苛养木杆菌菌株 9a5c 的寄主柑橘不同，拟南芥和番茄都有可追溯建立得很好的遗传系统，这为 GMI1000 基因组序列研究，提供了全面剖析植物与病原菌相互作用的借鉴。GMI1000 基因组包括一个 3.72Mb 染色体和一个 2.1Mb 大质粒[19]，染色体和大质粒的 G+C 百分比相近(分别为67%和66.9%)，这两个分子每百万碱基对编码的基因数相当，染色体编码 3448 个蛋白质，大质粒编码 1681 个蛋白质。已推测出共 2261（44.1%）个蛋白质的功能，有趣的是，染色体和大质粒上蛋白质的生物学功能有所不同，能够推测出功能的基因多位于染色体上，而大质粒中假定基因的比例高。但是，在毒性机制方面，大质粒编码一些涉及毒理的关键基因，如 *hrp* 基因（它是在植物中寄生所必需的，并涉及效应物分子分泌）和一些与产生胞外多糖产物有关的基因。

与苛养木杆菌不同的是，植物病理学家对青枯病菌研究很多，并且在基因组序列公布之前，病理学家已知道了该病原菌的一些致病机制[20]。然而，全基因组序列使病理学家能全面探索该病原菌的毒理，该基因组序列的 195 个新基因被认为是可能的病理基

因，这些包括编码效应物分子基因，编码黏附和依附因子基因、编码降解酶类基因、编码毒素基因、编码氧化压力反应因子基因等[19]。

根癌土壤杆菌

根癌土壤杆菌（*Agrobacterium tumefaciens*）在最近30年被广泛研究，因为它能将DNA转移到植物细胞核中构建转基因植株。经过传统分子生物学研究，已解析了根癌土壤杆菌转移DNA的机制，并将其进一步应用于生物工程领域。

两个独立研究组各自分别对同一根癌土壤杆菌菌株进行了测序，表明完全理解生物学的重要性[22,23]。根癌土壤杆菌C58基因组由4个复制元组成：2.8Mb环状染色体、2.1Mb线状染色体、543kb质粒（pATC58）和214kb质粒（pTiC58）[23]，基因组共编码5419个蛋白质，其中大部分（64.1%）已推测出其功能。复制子的G+C百分比相似，从56.7%～59.4%，然而，该菌向真核植物寄主细胞核转移DNA（T-DNA）的GC百分比却很低，pTi58质粒T-DNA的GC百分比是46%[23]。

在获得基因组序列数据前，根癌土壤杆菌毒理的大量信息已获得，或许，根癌土壤杆菌基因组序列最有兴趣的结果是，根癌土壤杆菌与两种植物共生菌的高度相似性，它们是苜蓿根瘤菌（*Sinorhizobium meliloti*）和百脉根瘤菌（*Mesorhizobium loti*）[25]，与百脉根瘤菌相比，苜蓿根瘤菌更接近根癌土壤杆菌[23]。苜蓿根瘤菌染色体和百脉根瘤菌都是在植物根上形成团块的共生菌，可以固氮。将根癌土壤杆菌C58环状染色体与苜蓿根瘤菌染色体排在一起，揭示了二者高度共线性[22,23]。因此，虽然一个是有益菌，一个是致病菌，但是根癌土壤杆菌和苜蓿根瘤菌都是来自于土壤，并与植物相关。基因和基因序列的保守性证明了这一共性（见第12章）。

黄单胞杆菌

野油菜黄单胞杆菌野油菜致病变种（*Xanthomonas campestris* pv *campestris*）导致十字花科植物变黑腐烂[26]，并与青枯菌类似，属维管病原菌。变黑腐烂病是全世界范围内十字花科植物的重大病害[26]，所有商品生产的这类植物都容易受该病原菌侵袭，对该病害的控制很复杂，这是由于该病原菌由种子携带，而这类种子又缺乏抵抗力[26]。对十字花科植物拟南芥寄主的抗性机制进行了初步研究并有了一些认识，因为野油菜黄单胞杆菌野油菜致病变种也是这个模式植物品种的致病菌（对黄单胞杆菌与拟南芥相互作用的综述，参见文献［27］）。

地毯黄单胞菌柑橘致病变种（*Xanthomonas axonopodis* pv *citri*）导致柑橘腐烂，柑橘腐烂是全世界柑橘的重大病害，唯一的控制措施就是隔离和移走销毁感染的树木（有关该病的论文见［28］）。与野油菜致病变种相比，柑橘致病变种感染的是非维管组织，导致形成叶和果实的损伤和腐烂[29]。

在一篇经典文献中，巴西ONSA研究组报告了两个品种的基因组序列，即野油菜黄单胞杆菌野油菜致病变种ATCC33913和地毯黄单胞菌柑橘致病变种306[30]。这两种病原菌的基因组大小相似，前者有一个5.1Mb环状染色体，后者有一个5.2Mb环状染色体以及两个34kb和65kb小质粒，从两个种的基因组中分别识别了4182个和4313个可读框，出现了高度保守的基因序列和基因排列顺序，大约82%可读框在氨基酸水平

有80%以上的一致性。在共线性方面，来自一个基因组的大约70%非转座因子基因（nontransposable element gene）能够与另一基因组中相应的直系同源基因（orthologous gene）线性排列。每一基因组中都有少量品种专一性基因，野油菜致病变种有646个，柑橘致病变种有800个。

这两个黄单胞菌的毒理和病理机制高度保守，编码 *hrp* 调节子、效应物分子、蛋白酶和细胞壁降解酶基因都存在这两个基因组中[30]。然而，这两个病原菌在特定类型的毒理和病理因素的数量和代表性不同。例如，野油菜致病变种比柑橘致病变种有更多涉及细胞壁的降解基因，这表明这些附加降解基因的数目和特异性，导致了柑橘腐败和变黑霉烂不同症状的形成[30]。两种菌编码效应物分子类型的基因也不同，野油菜致病变种编码效应物分子的类型范围更广[30]。显然，对每一品种特定基因的鉴定，其中病理和毒理基因只是一个分支，提供了系统解剖细菌病原菌生长繁殖需求的良好资源，这些病原菌定居在两个不同的生态位（维管系统和叶肉组织）并导致两个不同的病害症状（腐烂和溃疡）。

植物病原菌基因组研究进展

有几个细菌性病原菌基因组计划正在进行中，随着这些基因组的完成，更广范围植物细菌性病原菌的基因序列即可利用。欧文氏菌属（*Erwinia*）的两个种的基因组序列正在进行中（表1），欧文氏菌属与大肠埃氏菌属相近。此外，革兰氏阳性第一个植物病原菌，密执安棒形杆菌（*Clavibacter michiganensis* spp. *sepedonicus*）正在英国桑格研究中心（Sanger Center）进行测序（表1）。正处于测序最后阶段的一个基因组是丁香假单胞菌番茄致病变种DC3000，它引起番茄细菌斑，其基因组大小约6.5Mb，包括6.4Mb染色体和两个约70kb质粒（C. R.Buell，未发表），即使没有完整基因组，但对未完成基因组的数据挖掘，已发现了该病原菌更多涉及病理效应物分子[31,32]。感染豆类植物的第二个变种是丁香假单胞菌番茄致病变种B728a分离株，已被DOE-JGI测序到草图水平（表1），将为比较分析丁香假单胞菌变种群提供资源。

植物卵菌和真菌性基因组计划

由于细菌病原菌基因组较小，并且只有少数属的细菌感染植物，细菌性病原菌已在基因组的舞台上受到最先和最大的关注，而卵菌、真菌或线虫植物病原并没有。相比而言，代表分类多样性的大量真菌和卵菌能感染植物，许多属和种的真菌或卵菌病原菌基因组序列需要测定，而种的平均基因组大小比细菌大一个数量级，因此，需要比细菌基因组测序计划多得多的经费。

真菌许多种对植物是致病的，包括子囊菌纲、担子菌纲和不完全真菌，不完全真菌失去了有性阶段，只通过无性方式繁殖。卵菌菌纲，过去属于真菌界，现归于 *Stramenopile* 界[33,34]。美国植物病理学会（American Phytopathological Society）试图缩小基因组测序的范围，列出了优先测序的病原菌（http://www.apsnet.org/media/ps/top.asp），这是根据病害的经济相关性、科学团体的兴趣、遗传可溯性和其他影响测序信息的工具和资源。在该表中，美国植物病理学会为基因组测序推荐了26个真菌或卵

菌，基因组大小从玉米黑穗病菌（*Ustilago maydis*）的20Mb到马铃薯或番茄晚期枯萎病菌（*Pytophthora infestans*）的237Mb。

通过表达序列标签，可以获得对真菌基因组编码区的认识，其中，从环状DNA的克隆，可获得单一序列（single-pass sequence），建立特殊发育或生物周期cDNA文库，得到不同病理、毒理或繁殖阶段的转录组（transcriptome）样本。至今，真菌或卵菌许多植物病原菌的表达序列标签计划正在进行中，在基因库中共有31 985个表达序列标签（截至2002年5月10日）。表2列出了真菌或卵菌病原菌、病害及每个种的表达序列标签数目，虽然表达序列标签很少，却代表进入真菌和卵菌植物病原菌基因组的第一步。

植物病原线虫基因组计划

植物寄生线虫是农业病害损失的重要组成部分（对植物病原线虫的综述，参见文献[35]）。作为土壤病原的存在，不可简单地通过烟熏土壤来控制。两个属的根结线虫（*Meloidogyne*）和包囊线虫（*Heterodera*），在基因组学方面受到广泛关注。与真菌和卵菌植物病原菌一样，植物寄生线虫的基因组太大，无法用当前的经费进行全面基因组测序，因此，表达序列标签就成为植物寄生线虫基因组计划的起始点（表3）。

除了通过表达序列标签计划来发现新基因外，预期对植物寄生线虫基因组的主要认识，将通过把它们的基因组表达序列标签与模式线虫秀丽线虫（*Caenorhabditis elegans*）的基因组全序列（已被完全测序[36]）进行比较来获得。除了秀丽线虫已被测序外，另一个基因组计划将为植物寄生线虫提供第二个基因组（http://www.nematode .net/Species. Summaries/Caenorhabditis. briggsae）。

结论

植物病理学家们正通过已经完成或正在进行的细菌性植物病原菌基因组，第一次品尝到基因组的滋味。利用此资源将会对该领域产生深远的影响——相当于克隆第一个病原菌无毒基因和互补寄主抗性基因，因此，证实了40年前提出的基因对基因假说。病理学家们不是一次测定一个基因，而是将进行基因组范围内的分析，揭示涉及病理和毒理的所有方面。根据这些研究，将开发新的控制机制来发展有效高产农业。此外，早期用于植物病原菌的比较基因组学，提供涉及相关菌株、变种、种或属的寄主特异性基因识别手段，将继续是植物病原菌基因组研究的有力工具。

致谢

对丁香假单胞菌番茄致病变种的研究，是由国家科学基金植物基因组研究项目（DBI0077622）资助。作者们非常感激 Vinita Joardar 对本文的严格审阅，感谢 Ama Kwamena-Poh 给予管理上的支持。

（汪世山　译）

参考文献

1. Flor HH. Current status of the gene-for-gene concept. Annu Rev Phytopathol 1971; 9:275–296.
2. Staskawicz BJ. Genetics of plant-pathogen interactions specifying plant disease resistance. Plant Physiol 2001; 125:73–76.
3. Gopalan S, Bauer DW, Alfano JR, Loniello AO, He SY, Collmer A. Expression of the *Pseudomonas syringae* avirulence protein AvrB in plant cells alleviates its dependence on the hypersensitive response and pathogenicity (Hrp) secretion system in eliciting genotype-specific hypersensitive cell death. Plant Cell 1996; 8:1095–1105.
4. Thilmony RL, Chen Z, Bressan RA, Martin GB. Expression of tomato *Pto* gene in tobacco enhances resistance to *Pseudomonas syringae* pv *tabaci* expressing *avrPto*. Plant Cell 1995; 7: 1529–1536.
5. Rommens CMT, Salmeron JM, Oldroyd GED, Staskawicz BJ. Intergeneric transfer and functional expression of the tomato disease resistance gene *Pto*. Plant Cell 1995; 7:1537–1544.
6. Whitham S, McCormick S, Baker B. The *N* gene of tobacco confers resistance to tobacco mosaic virus in transgenic tomato. Proc Natl Acad Sci USA 1996; 93:8776–8781.
7. Hammond-Kosack KE, Tang S, HarrisonK, Jones JDG. The tomato *Cf-9* disease resistance gene functions in tobacco and potato to confer responsiveness to the fungal avirulence gene product Avr9. Plant Cell 1998; 10:1251–1266.
8. Simpson AJ, Reinach FC, Arruda P, et al. The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 2000; 406:151–157.
9. Wells JM, Raju BC, Hung H-Y, Weisburg WG, Mandelco-Paul L, Brenner DJ. *Xylella fastidiosa* gen nov, sp nov: Gram-negative, xylem-limited, fastidious plant bacteria related to *Xathomonas* species. Int J Syst Bacteriol 1987; 37:136–143.
10. Hopkins DL. *Xylella fastidiosa*: xylem-limited bacterial pathogens of plants. Annu Rev Phytopathol 1989; 27:271–290.
11. Li WB, Zreik L, Fernandes NG, et al. A triply cloned strain of *Xylella fastidiosa* multiplies and induces symptoms of citrus variegated chlorosis in sweet orange. Curr Microbiol 1999; 39: 106–108.
12. Derrick KS, Timmer LW. Citrus blight and other diseases of recalcitrant etiology. Annu Rev Phytopathol 2000; 38:181–205.
13. Alfano JR, Collmer A. The type III secretion pathway of plant pathogenic bacteria: trafficking harpins, avr proteins, and death. J Bacteriol 1997; 179:5655–5662.
14. He SY. Type III protein secretion systems in plant and animal pathogenic bacteria. Annu Rev Phytopathol 1998; 36:363–392.
15. Staskawicz B, Mudgett MB, Dangl J, Galan JE. Common and contrasting themes of plant and animal diseases. Science 2001; 292:2285–2289.
16. Hendson M, Purcell AH, Chen D, Smart C, Guilhabert M, Kirkpatrick B. Genetic diversity of Pierce's disease strains and other pathotypès of *Xylella fastidiosa*. Appl Environ Microbiol 2001; 67:895–903.
17. Schell MA, Denny TP, Huang J. Extracellular virulence factors of *Pseudomonas solanacearum*: role in disease and their regulation. In: Kado CI, Crosa JH (eds). Molecular Mechanisms of Bacterial Virulence. Dordrecht, The Netherlands: Kluwer, 1994, pp. 311–324.
18. Deslandes L, Pileur F, Liaubet L, et al. Genetic characterization of *RRS1*, a recessive locus in *Arabidopsis thaliana* that confers resistance to the bacterial soilborne pathogen *Ralstonia solanacearum*. Mol Plant Microbe Interact 1998; 11:659–667.
19. Salanoubat M, Genin S, Artiguenave F, et al. Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature 2002; 415:497–502.

20. Schell MA. Control of virulence and pathogenicity genes of *Ralstonia solanacearum* by an elaborate sensory network. Annu Rev Phytopathol 2000; 38:263–292.
21. Zupan J, Muth TR, Draper O, Zambryski P. The transfer of DNA from *Agrobacterium tumefaciens* into plants: a feast of fundamental insights. Plant J 2000; 23:11–28.
22. Goodner B, Hinkle G, Gattung S, et al. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. Science 2001; 294:2323–2328.
23. Wood DW, Setubal JC, Kaul R, et al. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science 2001; 294:2317–2323.
24. Galibert F, Finan TM, Long SR, et al. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 2001; 293:668–672.
25. Kaneko T, Nakamura Y, Sato S, et al. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res 2000; 7:331–338.
26. Williams PH. Black rot: a continuing threat to world crucifers. Plant Disease 1980; 64:736–742.
27. Buell CR. Interactions of *Arabidopsis* with *Xanthomonas*. In: Somerville CR, Meyerowitz EM (eds). The Arabidopsis Book. Rockville, MD: American Society of Plant Biologists. 2002; DOI 10.119/tab.0031, http://www.aspb.org/publications/arabidopsis/.
28. Brown K. Florida fights to stop citrus canker. Science 2001; 292:2275–2276.
29. Gottwald TR, Graham JH. In: Timmer LW, Garsney SM, Graham JH (eds). Compendium of Citrus Diseases. St. Paul, MN: American Phytopathological Society, 2002, pp. 5–7.
30. Da Silva AC, Ferro JA, Reinach FC, et al. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature 2002; 417:459–463.
31. Fouts DE, Abramovitch RB, Alfano JR, et al. Genome-wide identification of *Pseudomonas syringae* pv *tomato* DC3000 promoters controlled by the HrpL alternative sigma factor. Proc Natl Acad Sci USA 2002; 99:2275–2280.
32. Petnicki-Ocwieja T, Schneider DJ, Tam VC, et al. Genomewide identification of multiple proteins secreted by the Hrp type III protein secretion system of *Pseudomonas syringae* pv *tomato* DC3000 on the basis of N-terminal export-associated patterns, Hrp promoters, and horizontal transfer indicators. Proc Natl Acad Sci USA 2002; 99:7652–7657.
33. Sogin ML, Morrison HG, Hinkle G, Silberman JD. Ancestral relationships of the major eukaryotic lineages. Microbiologia 1996; 12:17–28.
34. Sogin ML, Silberman JD. Evolution of the protists and protistan parasites from the perspective of molecular systematics. Int J Parasitol 1998; 28:11–20.
35. Bird DM, Kaloshian I. Are roots special? Nematodes have their say. Physiol Mol Plant Pathol, in press.
36. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 1998; 282:2012–2018.

14

不产氧光合细菌

F. Robert Tabita and Thomas E. Hanson

引言

许多原核微生物能利用光化学反应获取能量，这些微生物包括各种不产氧细菌（多形细菌、绿硫细菌、绿色非硫细菌）和一些产氧（蓝细菌、原绿藻）光合细菌。这些光合细菌将光化学能量产物与其他影响生物圈极其重要的过程耦联在一起。另外，嗜极端环境微生物，如嗜盐古生菌也能将光能转化成可利用的化学能，但是速度相当慢，范围非常有限，因此，在这里对光合原核生物的讨论中将不考虑古生菌。然而，关于光合多形细菌和蓝细菌的生理学、生物化学、分子生物学以及生态学等方面的信息量巨大，为基本了解这些生物复杂的生活模式及它们在分子水平的代谢调控方法，读者需要阅读20世纪90年代中期发表的两个完整专著，这些书能使研究者们深入了解不产氧光合菌[1]和蓝细菌[2]的生理学和分子生物学的各个方面。

近年来，发表了大量的研究，仅凭这一章篇幅想阐述所有已发生的内容是不可能的，但是，阐述基因组学怎样推动近期研究则是可行的。已经有光合基因组与非光合基因组间的比较，但更应该留意我们熟悉的那些生物：紫色非硫细菌沼泽红假单胞菌（*Rhodopseudomonas palustris*）、球形红杆菌（*Rhodobacter sphaeroides*）、荚膜红杆菌（*Rhodobacter capsulatus*）和深红红螺菌（*Rhodospirillum rubrum*）以及绿硫细菌微温杆状绿菌（*Chlorobium tepidum*），其基因序列完成图或草图很有用[3]。此外，将集中研究 CO_2 固定、固氮、硫氧化、氢代谢等基本过程，正是微生物的这些代谢能力而引起对它们的研究兴趣，我们将着重阐述新基因组序列为代谢控制的相关基础问题怎样指明新方向和新方法。

紫色非硫细菌是地球上发现新陈代谢机制最多样的生物，并成为探索许多重要生命过程的模式生物。这些细菌能采用微生物的5种主要新陈代谢模式：需氧和厌氧化学异养型、需氧化能自养型、厌氧光能自养型和光能异养型。沼泽红假单胞菌与其他紫色非硫细菌或任何现存的生物相比，更具独特能力，能够在单细胞中催化更多生化过程。沼泽红假单胞菌能以上述所有代谢模式生长，它在光能异养生长对有机碳源（包括芳香族碳水化合物和木质素单体）的利用，以及在以二氧化碳为唯一碳源进行光能自养对电子供体（还原硫化合物和氢气）利用，都有相当强的灵活性。也许是这种灵活性，沼泽红假单胞菌通常是用富集培养基分离时得到数量最大的紫色非硫细菌[4]，而且广泛存在土壤和水体中。毫无疑问，沼泽红假单胞菌在不同环境中都具有降解和循环地球上最丰富聚合物——木质组织单体的独特能力，使它得以广泛分布。

光合菌与非光合菌基因组的大致比较

已完成或正在进行的一系列光合细菌基因组测序工作如表1所示，它编译自美国国家生物技术信息中心数据库（NCBI）[5]、基因组在线数据库（Genomes OnLine Database）[6]和美国能源部联合基因组研究所（DOE-JGI）[7]。一目了然，与相当少数其他类群不产氧光合细菌的基因组计划相比，有关蓝细菌和原绿藻 prochlorophyte 的基因组计划占绝对优势，这种不相称要引起注意，因为许多独特和有生物学意义的现象，往往在被忽视的生理多元化一些不产氧的光合菌中发现。

表1 已完成和正在进行基因组测序计划的光合细菌

微生物	基因组大小/Mb	来源	进展状况
单细胞蓝细菌（Unicellular cyanobacteria）			
集胞藻属 sp PCC 680（*Synechocystis* sp PCC680）	33.57	Kazusa DNA 研究院	已发表[50,51]
聚球蓝菌属 sp PCC 6301（*Synechococcus* sp PCC 6301）	2.69	Nagoya 大学	正在测序
聚球蓝菌属 sp WH8102（*Synechococcus* sp WH8102）	2.72	DOE-JGI	正在注解
聚球蓝菌属 sp PCC7942（*Synechococcus* sp PCC7942）	2.5	Texas A&M 大学	正在测序
聚球蓝菌属 sp PCC7002（*Synechococcus* sp PCC7002）	3.2	北京大学	正在测序
好热性蓝细菌（*Thermosynechococcus elongatus*）BP1	2.6	Kazusa DNA 研究院	已发表[52]
丝状蓝细菌（*Filamentous cyanobacteria*）			
鱼腥蓝细菌属（*Anabaena* sp PCC7120）sp PCC7120	6.4+0.8	Kazusa DNA 研究院	已发表[53]
点型念珠蓝细菌（*Nostoc punctiforme*）ATCC 29133	9.76	DOE-JGI	正在注解
铜绿微囊蓝细菌（*Microcystis aeruginosa*）	3.15	巴斯德研究所	正在测序
原绿藻（Prochlorophytes）			
海洋原绿球藻（*Prochlorococcus marinus*）MED4	1.66	DOG-JGI	正在注解
海洋原绿球藻（*Prochlorococcus marinus*）MIT9313	2.4	DOG-JGI	正在注解
海洋原绿球藻（*Prochlorococcus marinus*）SS120	1.8	Genoscope	正在测序
紫色非硫细菌（Purple Nonsulfur）			
荚膜红杆菌 SB1003（*Rhodobacter capsulatus* SB1003	3.7	Integrated Genomics	已完成
沼泽红假单胞菌 CGA009（*Rhodopseudomonas palustris*）CGA009	5.47	DOG-JGI	待发表
球形红杆菌 2.4.1（*Rhodobacter sphaeroides* 2.4.1）	4.4	DOG-JGI	正在测序
深红红螺菌（*Rhodospirillum rubrum*）	3.4	DOG-JGI	正在注解
紫色硫细菌（Purple sulfur）			
好热性光合细菌（*Thermochromatium tepidum*）	3.30	Integrated Genomics	正在注解
螺旋菌（Heliobacteria）			
运动螺旋菌（*Heliobacillus mobilis*）	4.2	Integrated Genomics	已完成
绿色非硫细菌（Green Nonsulfur）			
橙色绿屈挠杆菌 J-10-f1（*Chloroflexus aurantiacus* J-10-f1）	3	DOE-JGI	正在测序
绿硫细菌（Green sulfur）			
微温杆状绿菌（*Chlorobium tepidum*）	2.2	TIGR	已发表[3]

用公开可利用的已完成序列图和序列草图注释，将光合细菌基因组（9个基因组、5个草图、4个完成图）与非光合细菌（22个基因组）和古生菌（9个基因组）代表类

群的基因组进行全面比较。一些基因组极度简并的微生物基因组序列，如支原体及其亲缘种[7~9]或一些严格细胞内病原体，将排除在比较之外。从这项比较中得出很明显的一种趋势，光合细菌基因组比其他细菌或古生菌基因组数据的平均值大（图 1)。这可能是倾向对细菌性病原体和嗜极端环境微生物测序，这些细菌的生理多样性有限，因而只需相对较小基因组编码必需的代谢途径，这种趋势在营养要求苛刻或严格共生或寄生的细菌中尤为明显，这些细菌通常基因组很小[10]。

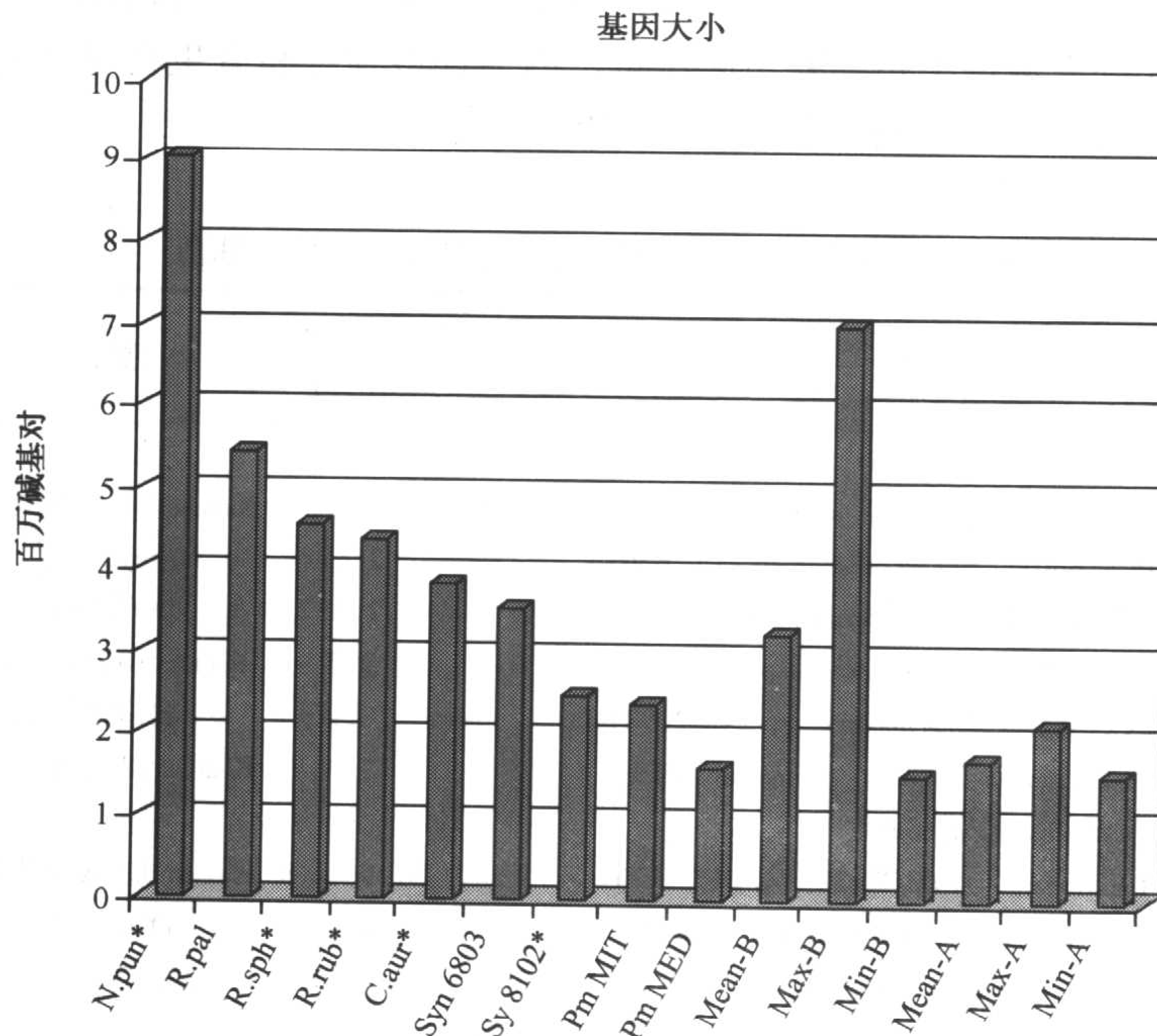


图 1　单个光合细菌基因组与 22 个非光合细菌基因组全序列（mean-B、Max-B 和 min-B）以及 9 个古生菌全序列（mean-A、Max-A 和 min-A）的平均值、最大值和最小值之间的比较。星号标记的光合细菌基因组表示草图序列［多于一个重叠群（contig)］，其他序列为一个完整的重叠群。*N. pun*，点型念珠蓝细菌；*R. pal*，沼泽红假单胞菌；*R. rub*，深红红螺菌；*C. aur*，橙色绿屈挠杆菌；*Syn* 6803，集胞藻 PCC6803；*Sy*8102，聚球蓝菌 WH8102；*Pm* MIT，*Prochlorococcus marinus* MIT9313；*Pm* MED，*Prochlorococcus marinus* MED4。非光合细菌基因组：风产液菌（*Aquifex aeolicus*)[54]，耐盐芽孢杆菌（*Bacillus halodurans*)[55]，枯草芽孢杆菌（*Bacillus subtilis*)[56]，空肠弯曲杆菌（*Campylobacter jejuni*)[57]，新月柄杆菌（*Caulobacter crescentus*)[58]，耐辐射异常球菌（*Deinococcus radiodurans*)[59]，大肠埃希氏菌 K12 和 O157:：H7（*Escherichia coli strains* K12[60] and O157：H7)[61]，流感嗜血菌（*Haemophilus influenzae*)[62]，幽门螺杆菌（*Helicobacter pylori*)[63]，乳酸乳球菌（*Lactococcus lactis*)[64]，百脉根中间根瘤菌（*Mesorhizobium loti*)[39]，结核分枝杆菌（*Mycobacterium tuberculosis*)[38]，麻风分枝杆菌（*Mycobacterium leprae*)[65]，脑膜炎奈瑟氏球菌 MC58 和 Z2491（*Neisseria meningiditis strains*) MC58[66] and Z2491[67]，多杀巴斯德氏菌（*Pasteurella multocida*)[68]，铜绿假单胞菌（*Pseudomonas aeruginosa*)[69]，酿脓链球菌（*Streptococcus pyogenes*)[70]，海栖热袍菌（*Thermotoga maritime*)[30]，霍乱弧菌（*Vibrio cholerae*)[71]和苛氧木杆菌（*Xylella fastidiosa*)[72]。古生菌基因组：敏捷气热菌（*Aeropyrum pernix*)[73]，闪烁古生球菌（*Archaeoglobus fulgidus*)[74]，盐杆菌（*Halobacterium* sp strain NRC-1)[75]，詹氏甲烷球菌（*Methanococcus jannaschii*）[76]，热自养甲烷杆菌（*Methanobacterium thermoautotrophicum*）[77]，极地古菌（*Pyrococcus horikoshii*）[78]，*Pyrococcus abyssi*，嗜酸热原体（*Thermoplasma acidophilum*)[79]，火山热原体（*T. volcanii*)[80]。

与直系同源群簇数据库（COG）中定义的某一家族相关每个基因组的可读框（ORF）总百分数相比较，发现直系同源序列很可能是某一共同祖先序列或相同超蛋白家族成员的后代[11,12]（图2）。光合细菌基因组中在COG数据库中有很多未描述过的可读框，这类可读框在细菌中平均为25%，在古生菌中为23%。值得注意的是，在所有COG种类中，光合型细菌未被充分代表，部分原因是在COG数据库中，编码光化学光捕获复合物和反应中心的结构基因是光合细菌所特有的。然而，预测参与光合色素（叶绿素、菌绿素、类胡萝卜素）生物合成的可读框，与在非光合细菌基因组中发现的蛋白质有很高的同源性，这种比较的结果表明，光合细菌基因组含有生理学潜能未被描述的一个遗传库。功能基因组学和大量综合性研究，如由能源部发起的沼泽红假单胞菌和球形红杆菌的微生物细胞项目（Microbial Cell Project），旨在揭示这些生理学潜能以及如何开发和控制它们。

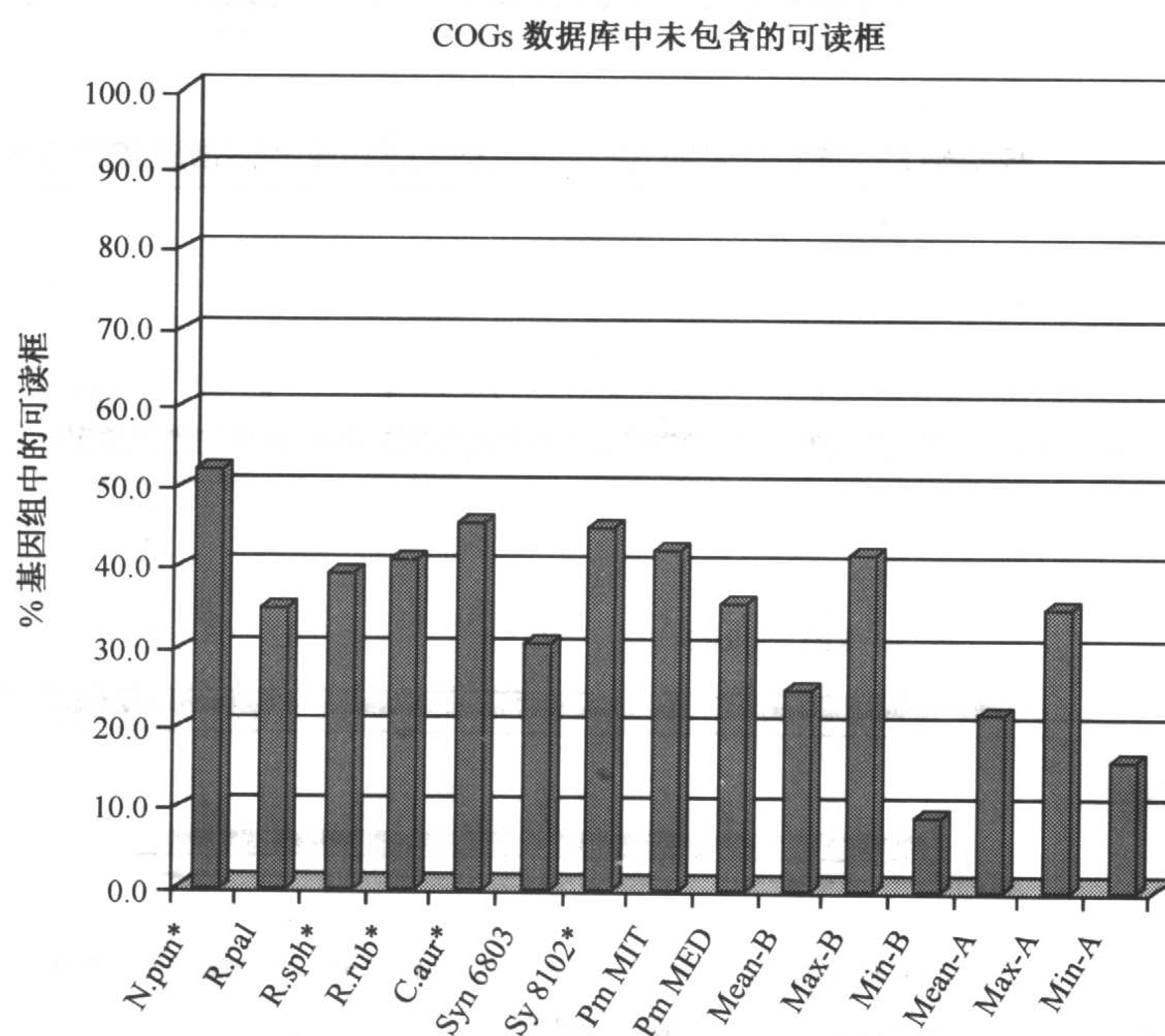


图2　对未列于直系同源群簇数据库(www.ncbi.nlm.nih.gov/COG)中的可读框占基因组序列百分数的比较。基因组缩写见图1。

二氧化碳的同化及其调控

紫色非硫细菌中二氧化碳的固定

所有紫色非硫细菌都是利用Calvin-Benson-Bassham（CBB）通道同化二氧化碳，绿硫细菌采用的是还原的三羧酸循环（TCA）[13~15]。在红杆菌属（*Rhodobacter*）和红假单胞菌属（*Rhodopseudomonas*）中，CBB基因安放在两个主要基因簇或操纵子中（cbb_{I} 和 cbb_{II}），每个都含有编码Ⅰ型或Ⅱ型核酮糖1,5二磷酸羧化酶/加氧酶（RubisCO）的基因，该酶催化真正的二氧化碳固定反应（图3）。在沼泽红假单胞菌和球形红

杆菌中，在 cbb_1 操纵子的第一个基因上游有一个反向转录 *cbbR* 基因，它编码的转录调节因子 CbbR 可激活两种 *ccb* 操纵子的转录[16]，这种情况在荚膜红杆菌中更加复杂，它的两个操纵子上游都有反向转录的调控基因，分别控制对应操纵子的转录[17]。

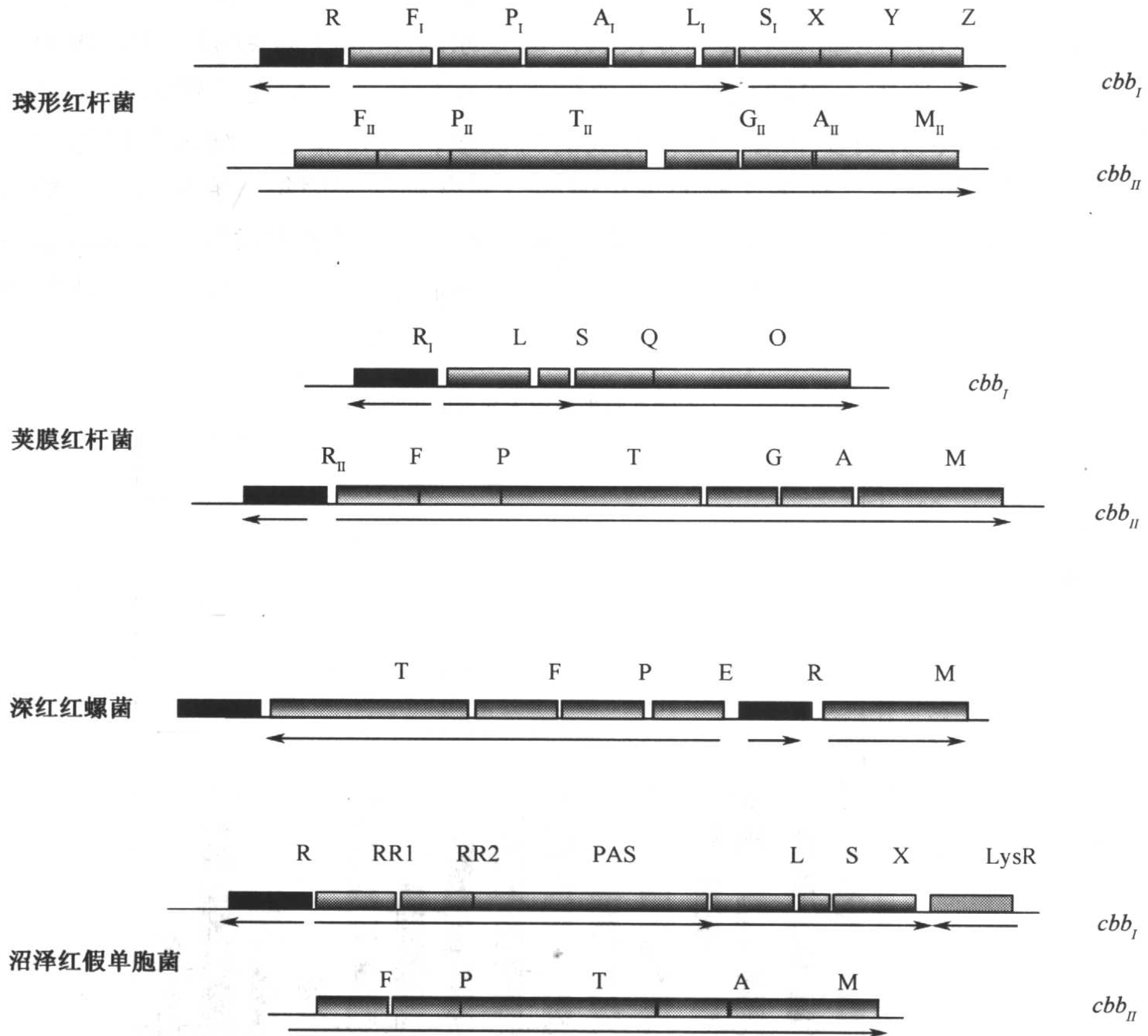


图 3　在紫色非硫细菌中二氧化碳固定结构基因（*cbb*）组成操纵子。在绿硫光合细菌（*C. tepidum*）中，还原性三羧酸循环的结构基因却不以基因簇的形式存在。

荚膜红杆菌 CbbRⅠ和 CbbRⅡ蛋白间的氨基酸序列相关性极低，比荚膜红杆菌 CbbRⅡ与其他微生物 CbbR 蛋白间的相关性还要低。事实上，荚膜红杆菌 *cbbR* 和 cbb_1 基因的比较序列分析表明，整个基因簇可能由某些化合自养型细菌经过水平基因迁移而获得[18]，毋庸置疑地造成了荚膜红杆菌的 CbbR 和 RubisCO 蛋白与球形红杆菌和沼泽红假单胞菌的不同。

在深红红螺菌中有一个主要 CBB 基因簇（图 3），值得注意的是，有三个有趣的可读框位于沼泽红假单胞菌的 *cbbR* 和 *cbbLS* 基因之间，这些基因编码（5′位于 *cbbL*，3′位于 *cbbR*，）两个推断的应答调节子（RR1 和 RR2）和一种大蛋白质（Pas），它包含几个有趣的模体（motif），其中有两个 PAS 结构域，一个 PAC 结构域，还有一个既是磷酸基团供体也是磷酸基团受体的结构域。最近研究表明，Pas 蛋白调节二氧化碳依赖型自养型细菌的生长和沼泽红假单胞菌 cbb_I 和 cbb_{II} 操纵子的转录（C.-S. Oh 和 F. R.

Tabita，未发表手稿)，这种独特的基因排列和大蛋白 Pas 的参与，启发我们对它们与 RR1 和 RR2 一起，在调节二氧化碳依赖性生长及其 CbbR 依赖性调控方面的机制研究。

在所有紫色非硫细菌中，那些控制光合体系生物合成[19~21]双组分调节系统（编码 RegA/B 或 PrrA/B 应答调节子-传感器激酶对）也参与调节 CO_2 的固定[19]。这些最新发现使我们注意到，在这些细菌及其相关细菌[22]（见下面）中，Reg（Prr）系统是控制和帮助整合几种重要生理过程的总调节因子。有充足证据表明，还有其他与二氧化碳固定有关的特异调节因子，*cbb* 操纵子在球形红杆菌和荚膜红杆菌中的调控明显不同[23]。

绿硫细菌中二氧化碳的固定

同紫色非硫细菌相比，绿硫细菌的代谢多功能性有限，因此，在光能自养和光能异养生长条件下，CO_2 固定能力和关键的 TCA 还原酶变化较小[24]。实际上，不像非硫细菌那样，在绿硫细菌中 CO_2 固定器必不可少，而且在微温杆状绿菌基因组中，没有迹象显示有其他生活方式的可能性。丙酮酸铁氧还原蛋白氧化还原酶/丙酮酸合成酶以及 α 酮戊二酸铁氧还原蛋白氧化还原酶/α 酮戊二酸合成酶，催化合成丙酮酸或 α 酮戊二酸的可逆反应，合成方式分别为氧化方式和还原方式。还原性羧化作用受还原性电子载体介导，因此，铁氧化还原蛋白是 CO_2 依赖性生长的关键因子，任何能引起这些电子载体还原的发生，都会影响细菌在 CO_2 固定过程中，是降解还是合成丙酮酸或 α 酮戊二酸[25,26]。

固氮能力

许多光养细菌有将氮气还原成氨的能力，可以在缺乏固定形式氮的环境中生长，这些特点使这些细菌在无机或有机氮缺乏的环境中有明显的选择优势。普通固氮酶复合体由两种主要成分组成，组分Ⅰ：二氮酶，包含黄素钼蛋白，它由两种不同多肽 NifD 和 NifK 组成；组分Ⅱ：二氮酶还原酶或称铁蛋白，由 NifH 多肽组成。在紫色非硫细菌中，*nifHDK* 基因为典型的共转录途径，由 *nifA* 基因产物调控。同时也存在对细胞氮和碳状态作出反应的复杂级连式氮调节，它牵涉到几种成分，包括 NtrA，NtrB，NtrC 和 *glnB*，*glnD* 基因的产物[27]。

传统上，在紫色非硫细菌中，*nif* 基因在氮气作为氮源或仅有少量有机氮源，如谷氨酸盐的生长状况下，通常是向上调节，最初在维涅兰德固氮菌（*Azotobacter vinelandii*）中发现另外两种固氮酶复合物，在这两种复合物中，组分Ⅰ或二氮酶中的辅助因子钼被钒或铁取代，相应的酶分别称为钒或铁固氮酶[28]，这些酶由 *vnfHDK* 或 *anfHDK* 基因编码。沼泽红假单胞菌，像维涅兰德固氮菌一样，是少数几种包括所有三种固氮酶的微生物，每种固氮酶都有相应的特异性调节因子，分别由 *nifA*、*vnfA* 和 *anfA* 基因编码。

每个系统的基因都有一种独特而有趣的组织方式，阐明每种固氮酶系统在代谢中的角色和作用非常重要。通过比较，荚膜红杆菌包括 *nif* 和 *anf* 系统，而球形红杆菌和微温杆状绿菌仅有 *nif* 系统（图 4)。深红红螺菌有 *nif* 和 *anf* 系统[29]，但是目前还没有

证明 *anf* 基因产物的功能。微温杆状绿菌 *nif* 操纵子的结构与产甲烷古生菌热自养甲烷杆菌的操纵子结构相似，暗示过去的遗传信息在某种情况下进行过跨域水平迁移。对微温杆状绿菌和其他基因组[30]的基因进行分析，也显示曾发生过比从前预期更频繁大量遗传信息的水平转移。

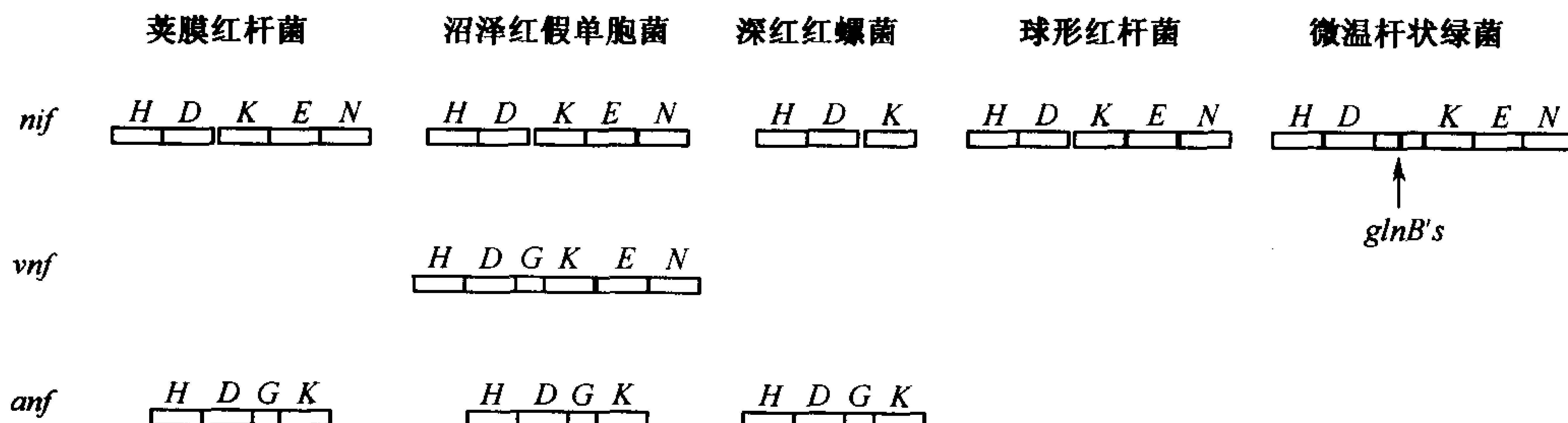


图 4 紫色非硫细菌和微温杆状绿菌（*C. tepidum*）固氮基因的聚类，包括含钼、钒和铁的二氮酶及其相关基因。

芳香烃化合物的降解

有少数光合微生物本来就有代谢芳香烃化合物的能力，其中主要是沼泽红假单胞菌，它能在有氧和厌氧光合条件下代谢安息香酸盐衍生物[31]，它有大量基因编码代谢这些反应所需的酶，有氧降解编码了四种独特裂解环途径，分别特异性针对 protocatechuate，homoprotocatechuate，2，5 二羟苯乙酸 homogentisate 和乙酸苯酯（phenylacetate），每条途径包括一种特异的加双氧酶，在基因组中还发现了另外 15 种，预计是单加氧酶、双加氧酶或者是 P450，暗示可能比目前已知更多的复合物能被转化。有一种单加氧酶可能与硫的获得有关，它位于一个操纵子中，该操纵子与恶臭假单胞菌（*Pseudomonas putida*）S313 的脂肪磺酸酯代谢有关的操纵子非常相似。

在厌氧条件下，芳香族化合物通过已经研究得很透彻的途径——苯（甲）酰辅酶 A（C_OA）降解途径降解，其中利用了一种新还原酶[33]和类似 β 氧化途径的催化剂[34]。对沼泽红假单胞菌基因组的分析也表明，这种微生物在厌氧代谢过程中还能应用许多目前还不了解的酶，因为除了与苯甲酰 C_OA 降解途径有关酶外，它还编码 42 种 C_OA 连接酶和 8 个 β 氧化基因簇，推测这种遗传多态性保证了沼泽红假单胞菌更广泛利用常存在土壤中植物来源的芳香族复合物，使这种微生物在这样的环境中得以广泛分布。

二氧化碳固定、氮固定和氢代谢的综合调控

在紫色非硫细菌中，通过敲除（knock out）编码 CBB CO_2 固定途径的关键酶和独特酶基因，发现 CO_2 和 N_2 的固定以及 H_2 的代谢调节是相互联系的[22,35]，CBB 途径的结构基因主要在两个操纵子中，某些情况则由单个 *cbbR* 基因调控（图 3）。如果这种细菌完全在光养条件下生长，敲除两个 RubisCO 基因或者 PRK 基因（*cbbp*），将会导致某些有趣的适应性，当然，如果 CBB 循环途径受损微生物就不能光能自养生长。

正常情况下，在野生型菌株中，有机碳存在并被氧化时，紫色非硫细菌可以通过CBB途径用CO_2作电子受体，然而，CBB途径的缺失菌株行光异养生长时，在缺乏外来电子受体下，它必须寻找替代方法，消除因有机碳源氧化而产生的还原电子。我们发现这些细菌可以用不同（某种程度上是未知的）方法，通过其他氧还平衡机制解决这个问题。

其中非常特别的一种适应性很有趣，在氨存在情况下，许多不同紫色非硫细菌阻遏二氮酶复合物的合成，然后用该酶固有的氢化酶活性，还原质子产生H_2[22,35]，这样就帮助这些微生物“释放”还原电位；在那些利用CO_2作为优先电子受体的能力被削弱的菌株中，这些还原电位通常积累阻止细胞生长。因此，这些菌株会启动高耗能合成二氮酶复合物的途径和*nif*系统，而这些途径在正常情况下（有氨存在时）被抑制。

这些结果证明了另一个重要观点，即正常时CO_2、N_2和H_2的代谢被精密调节，同时三个过程的控制高度整和（图5）。现在，还不很清楚细菌在缺乏功能性CBB途径时，怎样感知什么时候启动和去阻遏*nif*系统，仅知道双组分RegA/B（PrrA/B）的调节系统与此有关[22]。目前倾向于用现代功能基因组为基础的方法，进行阐明去阻遏并加强气体产生的分子基础研究。

图5 紫色非硫光合细菌（NSP PS）中*cbb*基因表达信号传递和调节中各因子间相互作用模式图。上图显示CO_2（*cbb*）和氮（*nif*）调节系统之间的相互关系。最初信号细胞膜接收，它可以影响一些关键成分（?）的氧还电位，影响RegB/PrrB自身磷酸化和随后RegA-P（PrrA-P）的形成，因此，CbbR蛋白和RegA-P（PrrA-P）参与正调节。推测CbbR′通过在CO_2固定条件下或者其他有利于*cbb*转录条件下，结合一种生成的共诱导分子而被转变为活性形式CbbR。*glnB*基因表达受*cbb*系统调控[81]，同时，*glnB*又通过Ntr系统和NifA调控*nif*的去阻遏。CBB通路的封锁导致产生H_2，它是通过去阻遏产生固氮酶复合物氢化酶活性而生成[22]，固氮酶复合物及其固有的氢化酶活性，消除了菌株中因不能用CO_2作为电子受体而产生多余的电位不平衡。图中的P表示正调节状态下被激活的启动子-调节子区域（经许可，选自参考文献[82]）。

沼泽红假单胞菌和微温杆状绿菌中硫磺代谢的比较分析

微温杆状绿菌是一种绿硫细菌，沼泽红假单胞菌是少数几种能够氧化还原硫复合物，并用氧化产生的能量支持依赖 CO_2 的光能自养生长的紫色非硫细菌之一，对这两种基因组中潜在硫氧化代谢途径加以分析比较。微温杆状绿菌基因组序列[3]从基因组研究所（TIGR）(www.tigr.org)获得，而沼泽红假单胞菌序列从联合基因组研究所(www.jgi.doe.gov)获得。

微温杆状绿菌和沼泽红假单胞菌在生理学上完全不同。前者是一种专性光合厌氧微生物，能氧化硫，有一定同化碳、氮和硫的能力[36]；后者是一种兼性光合厌氧微生物，能氧化硫，有较强同化多种化合物并使之转化为细胞物质。从这些不同可以预计，而实际上也是，沼泽红假单胞菌的基因组长度是微温杆状绿菌的两倍多，前者为 5.45Mb，而后者仅为 2.21Mb，此外，这两种细菌都可以利用还原性硫化合物进行光合生长。

两种微生物全基因序列和遗传工具的获得，使它们成为非常有用的模式系统，帮助理解光合情况下的硫代谢，从微温杆状绿菌和沼泽红假单胞菌基因组序列的分析可以推测硫代谢的过程，见图 6，现在还有很多悬而未决的问题，其中有些可在下面讨论。在这些细菌中仅有很少生化信息有关硫代谢，显然基因组为进一步研究提供了更广阔的空间，特别是碳同化和硫氧化的综合调控，然而，完全基因组序列能提供更多假说，指导以后有关硫代谢的实验。

硫代谢基因的复制

沼泽红假单胞菌基因组显示，它具有广泛利用氧化态含硫化合物的能力，相反，微温杆状绿菌基因组编码很有限利用氧化态含硫资源的能力，这与它的含硫化物热温泉生境一致[36]。还原性硫酸盐的利用是微温杆状绿菌得以维持或生存的一种策略，因为还原态硫化合物是光合生长所必需[36,37]，两种细菌基因组都包含与硫代谢有关酶的遗传冗余。

沼泽红假单胞菌的冗余出现在硫化物的固定，O-乙酰丝氨酸硫化氢解酶（OAS）固定硫化物生成半胱氨酸，O-乙酰同型丝氨酸硫化氢解酶（OAHS）固定硫化物生成同型半胱氨酸，同型半胱氨酸甲基化生成甲硫氨酸。微温杆状绿菌仅为每种活性编码一种酶的类似物，沼泽红假单胞菌则编码两种有 OAS 活性的同源物和四种有 OAHS 活性的同源物，与其他 OAHS 蛋白系列相比，其中四种有 OAHS 活性的同源物彼此类似，并与沼泽红假单胞菌基因组中的 *nifE*、*fixLJ* 和 *fixK2* 或 *vnfNE* 基因紧密相关（T. E. Hanson, 和 F. R. Tabita, 未公开发表的结果)。这种联系的重要性现在还不清楚，但基因的相互联系预示，它们与固氮酶辅助因子生物合成（*nifE* 和 *vnfNE*）和调节（*fixLJ* 和 *fixK2*）相关，像上面提到的那样，这些可以暗示硫和氮代谢间存在整体联系。其他四种同源物与 OAHS 蛋白功能特征密切相似，如此大量的潜在 OAHS 活性还没有在近缘微生物基因组中发现，如苜蓿中华根瘤菌和百脉根中生根瘤菌，它们仅编码两种潜在的 OAHS 活性物[38,39]。由此可以推测，沼泽红假单胞菌有更多的途径进行硫磺或低分子质量硫醇的代谢。

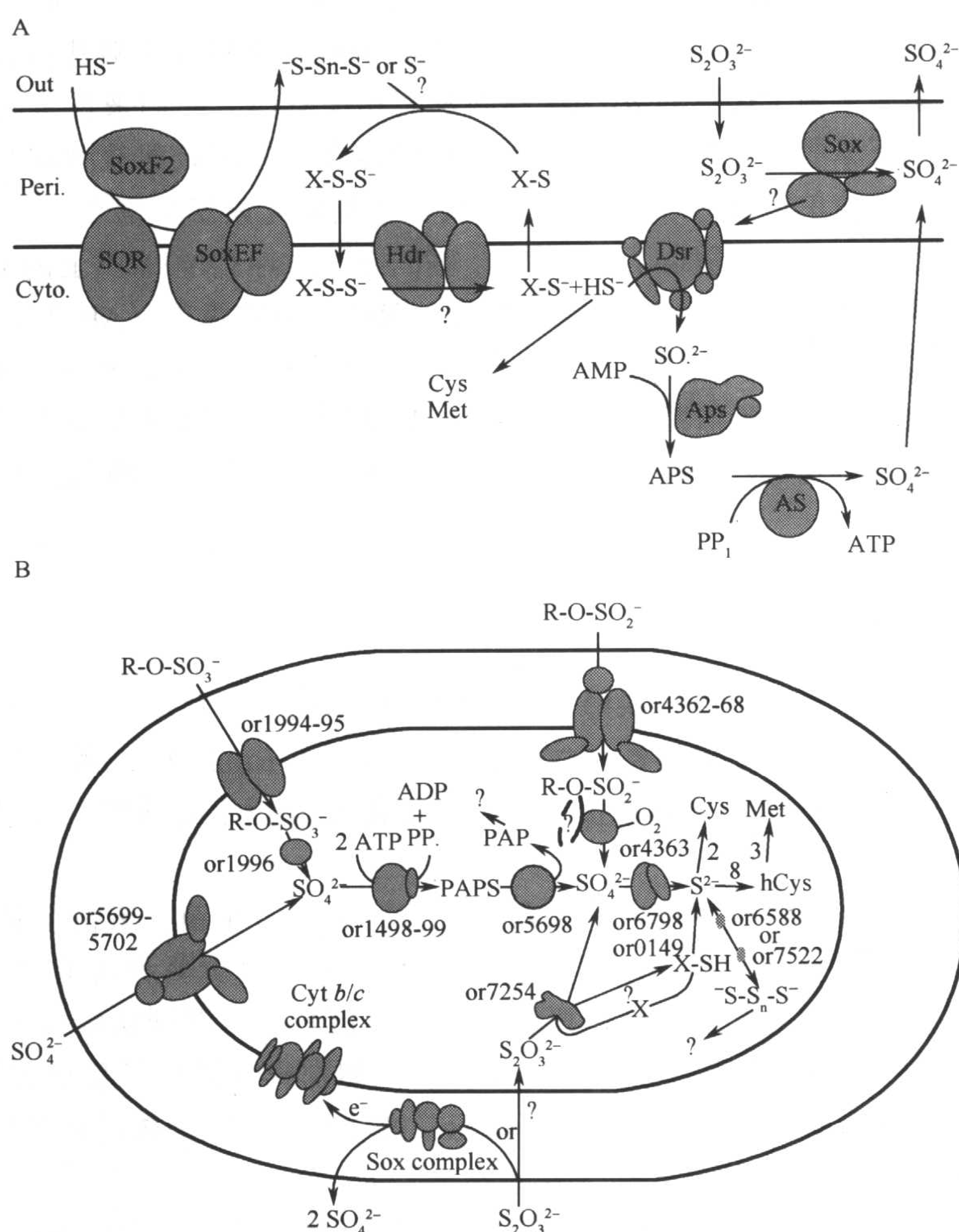


图 6　光合细菌中硫代谢途径。图中问号标记的问题需进一步实验证明。A. 微温杆状绿菌 *Chlorobium tepidum*。在已发表对该菌基因组系列进行分析时，已经提出了相似途径[3]。HS^-，硫化物；$^-S\text{-}S_n\text{-}S^-$，多聚硫化物；$S^0$，硫元素；$S_2O_3^{2-}$，硫代硫酸盐；$SO_3^{2-}$，亚硫酸盐；$SO_4^{2-}$，硫酸盐；X-$S^-$，低分子质量硫醇；SQR，硫化醌氧还酶；Sox，硫代硫酸盐氧化复合物；Hdr，可能是异二硫化物还原酶；Dsr，异化亚硫酸盐还原酶类似物；Aps，硫酸磷酸腺苷还原酶；AS，ATP-硫酸化酶。B. 沼泽红假单胞菌 *R. palustris*. 图中注释的“or”数字表示特异性可读框（如 or1994-95）。Cys（半胱氨酸）和 Met（甲硫氨酸）合成途径旁，标记数字表示每一步由基因组编码的直向同源酶的数目。图中简写注释：无机硫化物见图 6A；R-O-SO_3^{2-}，烃基或芳基硫酸盐；R-O-SO_2^-，烃基或芳基磺酸盐；PAPS，磷酸腺苷磷酸硫酸；PAP，磷酸腺苷磷酸盐；hCys，同型半胱氨酸。

微温杆状绿菌冗余蛋白与硫化物和亚硫酸盐之间的转化有关，还原性同化生成甲硫氨酸或半胱氨酸之前，亚硫酸盐首先被具同化力的亚硫酸盐还原酶还原成硫化物。沼泽红假单胞菌编码 CysI 和 CysJ 蛋白类似物，它们分别具有同化力的亚硫酸盐还原酶的 α

和 β 亚基，CysI 蛋白与最近才确定其特性的铜绿假单胞菌的 CysI 非常相似[40]。微温杆状绿菌没有 CysI 和 CysJ 蛋白类似物，与它不能还原性地利用亚硫酸盐的特性一致，然而，该菌基因组中有两个拷贝的 *dsrAB* 基因及其相关基因，它们来自紫硫光合细菌编码异化亚硫酸盐还原酶的基因[41]，类似的基因也存在能还原硫酸盐的细菌中。遗传研究表明，紫色硫细菌单拷贝的 *dsr* 操纵子是氧化存储硫元素所必需[41]，在微温杆状绿菌中的复制长度达 4.6Kb，覆盖了 *dsrCABL* 基因有一致序列 99% 以上的 DNA。同时，这些基因还编码三磷酸腺苷（ATP）硫酸化酶、硫酸化磷酸腺苷（APS）还原酶和异化二硫化物还原酶类似物，在该菌基因组中与 *dsr* 操纵子的一个拷贝紧密连锁（图 6）[3]。复制的生理学结果现在还不清楚，是否 *dsrCABL* 基因的两个拷贝都表达还是生长所必需，有个拷贝与 ATP-硫酸化酶不连锁，在 *dsrB* 基因中缺失两个碱基，从而造成移码突变[3]。

RubisCO 类似蛋白和硫代谢

当网站上提供了部分微温杆状绿菌基因组序列草图后，发现了一种与 RubisCO 非常相似的蛋白质，它的许多重要活性位点残基与真正的 RubisCO 一致，该蛋白质序列现已测定，并分析了其特性[42]，该蛋白质没有催化 CO_2 固定的活性，所以称为 RubisCO 类似蛋白（RLP）。

其他几种微生物基因组序列也发现含有 RLP，包括古生菌、紫色非硫细菌沼泽红假单胞菌和深红红螺菌、紫色硫细菌（图 7）。事实上有至少两种可能是三种 RLP 亚类，当更多基因组序列完成后，这些亚类会越来越清楚。微温杆状绿菌的一个突变株（Ω∷RLP），缺乏 RLP，在光合条件下不能像野生型那样有效地氧化硫元素[42]，如果在生长介质中添加半胱氨酸则可弥补这一功能，说明缺陷型突变株可能存在低分子质量硫醇的代谢障碍。

在光合细菌中，已经预测低分子质量硫醇与硫元素氧化有关[43]，特别是当紫色硫细菌氧化储存硫元素时，谷胱甘肽氨基化合物以硫醇和过硫化物形式循环[44]。*Chlorobium limicola* 和微温杆状绿菌含有新的结构特异低分子质量硫醇[45]（T. E. Hanson 和 F. R. Tabita，未发表），推测在还原性含硫化合物氧化时，RubisCO 类似蛋白在这些化合物的合成或循环中起作用。从微温杆状绿菌基因组没有编码氧化硫代硫酸盐的全套系统这一事实，可推断出 RLP 蛋白的这种功能。

沼泽红假单胞菌和微温杆状绿菌基因组都含有编码与 *Paracoccus pantotrophus* GB17 中硫氧化（*sox*）系统类似的操纵子，该操纵子编码硫代硫酸盐：细胞色素 *c* 氧化还原酶。有篇综述描述了这个系统的组成、功能和对各种细菌中 *sox* 基因的分布[46]。在所有已知基因组系列中，沼泽红假单胞菌的 *sox* 操纵子与 *P. pantotrophus* 的操纵子最类似，微温杆状绿菌操纵子缺少几种基因，包括编码 SoxCD 成分的基因，SoxCD 是一种外周质含钼硫脱氢酶，体外完全氧化硫代硫酸盐所必需[47]。SoxCD 酶的底物是一种蛋白质结合的半胱氨酸-S-硫化物复合物，它与 *sox* 复合物 SoxYZ 亚基的 SoxY 结合[48]，反应产物是硫酸盐和 SoxY 中再生的半胱氨酸。那么，微温杆状绿菌既然没有这种活性，它是怎样完全氧化硫代硫酸盐呢？

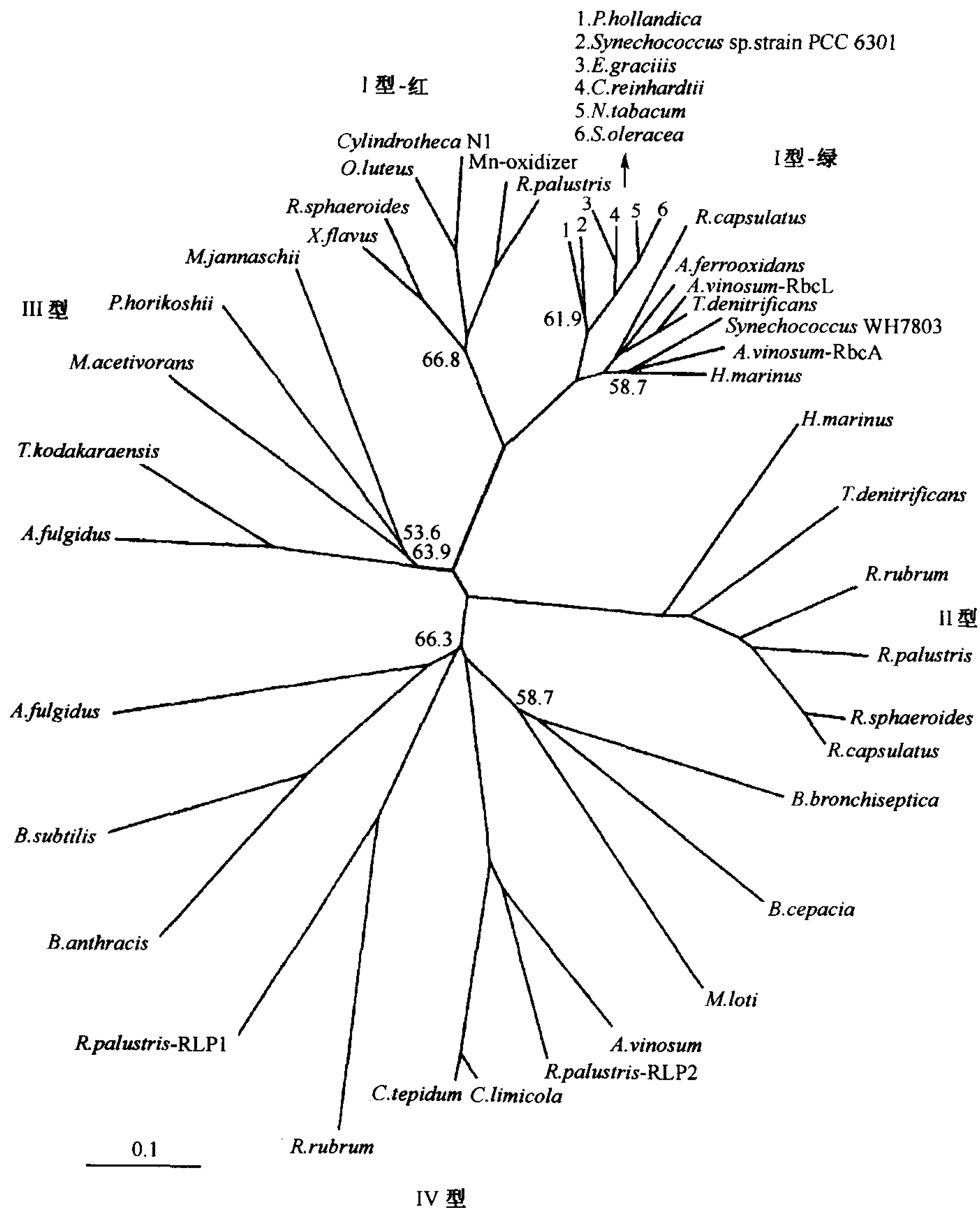


图 7 无根相邻种系发生树显示了真正 RubisCo 系列（I 型，红和绿；II 型，紫色；III 型，青铜色）与 RubisCO 类似蛋白系列（IV 型，蓝绿色）间的相互关系。节点位置的引导指令值显示，在 1000 次试验中特定节点出现的次数的百分比。仅标出小于 70% 引导指令的节点。刻度条代表每个位点替换值为 0.1。在文中其他地方未出现过的微生物全名包括：*A. ferroxidans*，*Acidithiobacillus ferroxidans*；*B. anthracis*，炭疽芽胞杆菌（*Baccilus anthracis*）；*B. cepacia*，洋葱伯克霍尔德氏菌（*Burkholderia cepacia*）；*B. sbutilis*，枯草芽孢杆菌（*Bacillus subtilis*）；*B. bronchiseptica*，支气管炎博德特氏菌（*Bordetella bronchiseptica*）；*C. reinhardtii*，莱茵衣藻（*Chlamydomonas reinhardtii*）；*E. gracilis*，纤细裸藻（*Euglena gracilis*）；*H. marinus*，海洋氢弧菌（*Hydrogenovibrio marinus*）；*M. acetivorans*，噬乙酸甲烷八叠球菌（*Methanosarcina acetivorans*）；*N. tabacum*，*Nicotiana tabacum*；*O. luteus*，*Olisthodiscus luteus*；*P. hollandica*，荷兰原绿丝蓝细菌（*Prochlorothrix hollandica*）；*S. oleracea*，*Spinacia oleracae*；*T. kodakaraenis*，超好热始原菌（*Thermococcus kodakaraensis*）；*T. denitrifcans*，脱氮硫杆菌（*Thiobacillus denitrificans*）；*X. flavus*，黄色黄杆菌（*Xanthobacter flavus*）。（另见文前彩色插图 14-7）

初步结果显示，温杆状绿菌突变株 Ω∷RLP 不仅氧化硫代硫酸盐有缺陷，而且氧化硫元素也有缺陷（T. E. Hanson 和 F. R. Tabita，未发表），突变株 Ω∷RLP 氧化硫化物并未受影响。我们推测，硫代硫酸盐氧化时，在 SoxY 上肯定存在一种半胱氨酸-S-硫化物的受体，该受体可能是低分子质量的硫醇。类似突变株 Ω∷RLP 中硫氧化的缺陷，我们假设由 RLP 缺失导致的低分子质量硫醇代谢缺陷，可解释突变株中硫代硫酸盐的氧化缺陷。SoxY 中硫代半胱氨酸的转化产物，应该是低分子质量的过硫化物，其结构类似甲烷生成和古生菌 C1 代谢中产生的异二硫化物。温杆状绿菌基因组编码异二硫化物氧还酶类似物，功能与 *Methanosarcina mazei* Göl[49]最接近。在温杆状绿菌基因组中，这些基因位于紫色硫细菌 *dsr* 系统（与硫元素氧化有关[41]）拷贝的基因附近，也与 ATP-硫酸化酶和 APS 还原酶的基因相邻。根据这些联系可以假设，硫元素和 SoxY 中硫代半胱氨酸的氧化，由低分子质量硫醇介导并经亚硫酸盐和 APS 生成硫酸盐，该假说的诱人之处是依靠 ATP-硫酸化酶的逆反应节约一个分子 ATP。

沼泽红假单胞菌也含有两种 RubisCO 类似蛋白，RLP1 和 RLP2，RLP2 与微温杆状绿菌的 RLP 蛋白在组成上有 66％氨基酸相同，RLP1 与该菌的 RLP 有 31％氨基酸相同。微温杆状绿菌的 RLP、沼泽红假单胞菌的 RLP2、*C. limicola* 和紫色硫细菌中的 RLP 序列一起形成 RLP 中的一个独特群簇（图 7），这些微生物都可在光合条件下氧化利用硫化物和硫代硫酸盐，RLP 在这些细菌中的特殊作用还不清楚，虽然，从沼泽红假单胞菌和温杆状绿菌的硫代谢基因比较分析，还不能得出一个具体的模型，但是，起码有了许多候选基因，可以对光养硫代谢进一步进行遗传和生化分析。

结论

光合细菌，包括其中一些可能是地球上代谢功能最广的微生物，它们可以在如此多样的条件下生存，催化许多基本的重要生化过程，并通过复杂手段调节代谢。最新测出的基因组使我们对这些代谢调节机制有了更深入的理解，同时，也为利用这些微生物进行固碳和替代能源的生产提供了研究依据。

致谢

在 Tabita 实验室进行的关于生物化学和二氧化碳同化以及对固氮、氢代谢、硫氧化的综合研究得到了能源部（DE-FG02-01ER63241 和 DE-FG02-91ER20033）和国立卫生研究院的支持（GM45404 和 GM24497）。

（程　萍　译）

参考文献

1. Blankenship RE, Madigan MT, Bauer CE, eds. Anoxygenic Photosynthetic Bacteria. Dordrecht, The Netherlands: Kluwer Academic Publishers,1995.
2. Bryant DA, ed. The Molecular Biology of Cyanobacteria. Dordrecht, The Netherlands: Kluwer, 1994.
3. Eisen JA, Nelson KE, Paulsen IT, et al. The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci USA 2002; 99: 9509–9514.
4. Gest H, Favinger JL. Enrichment of purple photosynthetic bacteria from earthworms. FEMS Microbiol Lett 1992; 91:265–270.
5. Wheeler DL, Church DM, Lash AE, et al. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 2002; 30:13–16.
6. Bernal A, Ear U, Kyrpides N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res 2001; 29:126–127.
7. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 2000; 407:757–762.
8. Chambaud I, Heilig R, Ferris S, et al. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res 2001; 29:2145–2153.
9. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 1996; 24: 4420–4449.
10. Gil R, Sabater-Munoz B, Latorre A, Silva FJ, Moya A. Extreme genome reduction in *Buchnera* spp: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci USA 2002; 99: 4454–4458.
11. Tatusov RL, Natale DA, Garkavtsev IV, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001; 29:22–28.
12. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 278:631–637.
13. Yoon KS, Hanson TE, Gibson JL, Tabita FR. Autotrophic CO_2 metabolism. In: Lederburg J (ed). Encyclopedia of Microbiology, 2nd ed. San Diego, CA: Academic, 2000, pp. 349–358.
14. Tabita FR. Molecular and cellular regulation of autotrophic carbon dioxide fixation in microorganisms. Microbiol Rev 1988; 52:155–189.
15. Buchanan BB, Arnon DI. A reverse KREBS cycle in photosynthesis: consensus at last. Photosynth Res 1990; 24:47–53.
16. Gibson JL, Tabita FR. The molecular regulation of the reductive pentose phosphate pathway in Proteobacteria and Cyanobacteria. Arch Microbiol 1996; 166:141–150.
17. Vichivanives P, Bird TH, Bauer CE, Tabita FR. Multiple regulators and their interactions in vivo and in vitro with the *cbb* regulons of *Rhodobacter capsulatus*. J Mol Biol 2000; 300:1079–1099.
18. Paoli GC, Soyer F, Shively J, Tabita FR. *Rhodobacter capsulatus* genes encoding form I ribulose-1,5-bisphosphate carboxylase/oxygenase (*cbbLS*) and neighbouring genes were acquired by a horizontal gene transfer. Microbiology 1998; 144:219–227.
19. Qian Y, Tabita FR. A global signal transduction system regulates aerobic and anaerobic CO_2 fixation in *Rhodobacter sphaeroides*. J Bacteriol 1996; 178:12–18.
20. Eraso JM, Kaplan S. *prrA*, a putative response regulator involved in oxygen regulation of pho-

tosynthesis gene expression in *Rhodobacter sphaeroides*. J Bacteriol 1994; 176:32–43.

21. Mosley CS, Suzuki JY, Bauer CE. Identification and molecular genetic characterization of a sensor kinase responsible for coordinately regulating light harvesting and reaction center gene expression in response to anaerobiosis. J Bacteriol 1995; 177:3359.
22. Joshi HM, Tabita FR. A global two component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation, and nitrogen fixation. Proc Natl Acad Sci USA 1996; 93:14,515–14,520.
23. Gibson JL, Dubbs JM, Tabita FR. Differential expression of the CO_2 fixation operons of *Rhodobacter sphaeroides* by the Prr/Reg two-component system during chemoautotrophic growth. J Bacteriol 2002; 184:6654–6664.
24. Wahlund TM, Tabita FR. The reductive tricarboxylic acid cycle of carbon dioxide assimilation: initial studies and purification of ATP-citrate lyase from the green sulfur bacterium *Chlorobium tepidum*. J Bacteriol 1997; 179:4859–4867.
25. Yoon KS, Hille R, Hemann C, Tabita FR. Rubredoxin from the green sulfur bacterium *Chlorobium tepidum* functions as an electron acceptor for pyruvate ferredoxin oxidoreductase. J Biol Chem 1999; 274:29,772–29,778.
26. Yoon KS, Bobst C, Hemann CF, Hille R, Tabita FR. Spectroscopic and functional properties of novel 2[4Fe-4S] cluster-containing ferredoxins from the green sulfur bacterium *Chlorobium tepidum*. J Biol Chem 2001; 276:44,027–44,036.
27. Kranz RG, Cullen PJ. Regulation of nitrogen fixation. In: Blankenship RE, Madigan MT, Bauer CE (eds). Anoxygenic Photosynthetic Bacteria. Dordrecht, The Netherlands: Kluwer, 1995, pp. 1181–1208.
28. Bishop PE, Premakumar R, Joerger RD, et al. Alternative nitrogen fixation systems in *Azotobacter vinelandii*. In: Bothe H, De Bruijn FJ, Newton WE (eds). Nitrogen Fixation: Hundred Years After. Stuttgart, Germany: Gustav Fisher, 1988, pp. 71–79.
29. Loveless TM, Bishop PE. Identification of genes unique to Mo-independent nitrogenase systems in diverse diazotrophs. Can J Microbiol 1999; 45:312–317.
30. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 1999; 399:323–329.
31. Harwood CS, Gibson J. Anaerobic and aerobic metabolism of diverse aromatic compounds by the photosynthetic bacterium *Rhodopseudomonas palustris*. Appl Environ Microbiol 1988; 54:712–717.
32. Kahnert A, Vermeij P, Wietek C, James P, Leisinger T, Kertesz MA. The ssu locus plays a key role in organosulfur metabolism in *Pseudomonas putida* S-313. J Bacteriol 2000; 182:2869–2878.
33. Gibson J, Dispensa M, Harwood CS. 4-Hydroxybenzoyl coenzyme A reductase (dehydroxylating) is required for anaerobic degradation of 4-hydroxybenzoate by *Rhodopseudomonas palustris* and shares features with molybdenum-containing hydroxylases. J Bacteriol 1997; 179:634–642.
34. Egland PG, Pelletier DA, Dispensa M, Gibson J, Harwood CS. A cluster of bacterial genes for anaerobic benzene ring biodegradation. Proc Natl Acad Sci USA 1997; 94:6484–6489.
35. Tichi MA, Tabita FR. Maintenance and control of redox poise in *Rhodobacter capsulatus* strains deficient in the Calvin–Benson–Bassham pathway. Arch Microbiol 2000; 174:322–333.
36. Wahlund TM, Woese CR, Castenholz RW, Madigan MT. A thermophilic green sulfur bacterium from New Zealand hot springs, *Chlorobium tepidum* Sp-Nov. Arch Microbiol 1991; 156: 81–90.
37. Mukhopadhyay B, Johnson E, Ascano M. Conditions for vigorous growth on sulfide and reactor-scale cultivation protocols for the thermophilic green sulfur bacterium *Chlorobium tepidum*. Appl Environ Microbiol 1999; 301–306.
38. Galibert F, Finan TM, Long SR, et al. The composite genome of the legume symbiont *Sinorhizo-*

bium meliloti. Science 2001; 293:668–672.

39. Kaneko T, Nakamura Y, Sato S, et al. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res 2000; 7:331–338.
40. Hummerjohann J, Kuttel E, Quadroni M, Ragaller J, Leisinger T, Kertesz MA. Regulation of the sulfate starvation response in *Pseudomonas aeruginosa*: role of cysteine biosynthetic intermediates. Microbiology 1998; 144:1375–1386.
41. Pott AS, Dahl C. Sirohaem sulfite reductase and other proteins encoded by genes at the *dsr* locus of *Chromatium vinosum* are involved in the oxidation of intracellular sulfur. Microbiology 1998; 144:1881–1894.
42. Hanson TE, Tabita FR. A ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO)-like protein from *Chlorobium tepidum* that is involved with sulfur metabolism and the response to oxidative stress. Proc Natl Acad Sci USA 2001; 98:4397–4402.
43. Brune D. Sulfur oxidation by phototrophic bacteria. Biochim Biophys Acta 1989; 975:189–221.
44. Bartsch R, Newton G, Sherril C, Fahey R. Glutathione amide and its perthiol in anaerobic sulfur bacteria. J Bacteriol 1996; 178:4742–4746.
45. Fahey RC, Buschbacher RM, Newton GL. The evolution of glutathione metabolism in phototrophic microorganisms. J Mol Evol 1987; 25:81–88.
46. Friedrich CG, Rother D, Bardischewsky F, Quentmeier A, Fischer J. Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? Appl Environ Microbiol 2001; 67:2873–2882.
47. Friedrich CG, Quentmeier A, Bardischewsky F, et al. Novel genes coding for lithotrophic sulfur oxidation of *Paracoccus pantotrophus* GB17. J Bacteriol 2000; 182:4677–4687.
48. Quentmeier A, Friedrich CG. The cysteine residue of the SoxY protein as the active site of protein-bound sulfur oxidation of *Paracoccus pantotrophus* GB17. FEBS Lett 2001; 503:168–172.
49. Ide T, Baumer S, Deppenmeier U. Energy conservation by the H_2:heterodisulfide oxidoreductase from *Methanosarcina mazei* Go1: identification of two proton-translocating segments. J Bacteriol 1999; 181:4076–4080.
50. Kaneko T, Sato S, Kotani H, et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions [supplement]. DNA Res 1996; 3:185–209.
51. Kaneko T, Sato S, Kotani H, et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 1996; 3:109–136.
52. Nakamura Y, Kaneko T, Sato S, et al. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. DNA Res 2002; 9:123–130.
53. Kaneko T, Nakamura Y, Wolk CP, et al. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp strain PCC 7120. DNA Res 2001; 8:205–213; 227–253.
54. Deckert G, Warren PV, Gaasterland T, et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 1998; 392:353–358.
55. Takami H, Nakasone K, Takaki Y, et al. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 2000; 28:4317–4331.
56. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 1997; 390:249–256.
57. Parkhill J, Wren BW, Mungall K, et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 2000; 403:665–668.
58. Nierman WC, Feldblyum TV, Laub MT, et al. Complete genome sequence of *Caulobacter crescentus*. Proc Natl Acad Sci USA 2001; 98:4136–4141.

59. White O, Eisen JA, Heidelberg JF, et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 1999; 286:1571–1577.
60. Blattner FR, Plunkett G 3rd, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997; 277:1453–1474.
61. Perna NT, Plunkett G 3rd, Burland GV, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 2001; 409:529–533.
62. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
63. Tomb JF, White O, Kerlavage AR, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997; 388:539–547.
64. Bolotin A, Wincker P, Mauger S, et al. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* sp *lactis* IL1403. Genome Res 2001; 11:731–753.
65. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001; 409:1007–1011.
66. Tettelin H, Saunders NJ, Heidelberg J, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
67. Parkhill J, Achtman M, James KD, et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 2000; 404:502–506.
68. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V. Complete genomic sequence of *Pasteurella multocida*, Pm70. Proc Natl Acad Sci USA 2001; 98:3460–3465.
69. Stover CK, Pham XQ, Erwin AL, et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 2000; 406:959–964.
70. Ferretti JJ, McShan WM, Ajdic D, et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 2001; 98:4658–4663.
71. Heidelberg JF, Eisen JA, Nelson WC, et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 2000; 406:477–483.
72. Simpson AJ, Reinach FC, Arruda P, et al. The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature 2000; 406:151–157.
73. Kawarabayasi Y, Hino Y, Horikawa H, et al. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 1999; 6:83–101, 145–152.
74. Klenk HP, Clayton RA, Tomb JF, et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 1997; 390:364–370.
75. Ng WV, Kennedy SP, Mahairas GG, et al. Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci USA 2000; 97:12,176–12,181.
76. Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 1996; 273:1058–1073.
77. Smith DR, Doucette-Stamm LA, Deloughery C, et al. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 1997; 179:7135–7155.
78. Kawarabayasi Y, Sawada M, Horikawa H, et al. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 1998; 5:55–76.
79. Ruepp A, Graml W, Santos-Martinez ML, et al. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature 2000; 407:508–513.
80. Kawashima T, Amano N, Koike H, et al. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc Natl Acad Sci USA 2000; 97:14,257–14,262.
81. Qian Y, Tabita FR. Expression of *glnB* and a *glnB*-like gene (*glnK*) in a ribulose bisphosphate carboxylase/oxygenase-deficient mutant of *Rhodobacter sphaeroides*. J Bacteriol 1998; 180:

4644–4649.
82. Tabita FR. Microbial ribulose-1,5-bisphosphate carboxylase/oxygenase: a different perspective. Photosynth Res 1999; 60:1–28.

15
嗜热微生物基因组

Frank T. Robb

引言：极端环境中的生命

生命系统的正常环境概念是指37℃、pH7.0、50 mmol/L NaCl和正常大气压。对高等真核生物而言极端生活条件只有概念上的意义，微生物界对常态有十分不同的观念，微生物生命适应范围远远超过了我们熟悉的中温环境，它们在几乎任何地质环境中都能够生存。表1列出了一些生活在已知生物圈边缘的微生物。例如，微生物生长的极限温度是大于或等于113℃或者是低于0℃；压力能够为至少50mPa；pH值能接近0；以及盐浓度能最多到饱和浓度（相当于5mol/L NaCl）[1]。在许多情况下，超过一种极端环境条件存在；例如，许多嗜压微生物也能够在高于水的正常沸点温度之上生长。目前已知最高生长温度的微生物发烟火叶菌（*Pyrolobus fumarii*）就是这种情况[2]；它是从中大西洋底极深的火山口（3650m）中分离出的，它在高温灭菌1小时的情况下仍能够存活，能够在113℃和25mPa的大气压下生长[2]。

表1 极端温度条件下的微生物

极端条件分类	环境	生物	限制生长条件	参考文献
高温生长（超嗜热微生物）	海底火山口	发烟火叶菌（*Pyrolobus fumarii*）(A), *Methanopyrus kandleri* (A)	T_{max}=113℃ T_{max}=110℃	Blochl et al.(1997); Huber et al.(1989); Slesarev et al.(2002)
高温生存	土壤，生长培养基的污染物	*Moorella thermoaceticum*（芽胞）(B)	2h, 121℃, 15psi	Bryer et al.(2000)
低温（嗜冷性微生物）	雪，湖水，沉积物，冰	无数，例如，弧菌，节杆菌，假单胞菌(B)和产甲烷菌(A)	－17℃	Carpenter et al. (2000); Cavicchioli and Thomas (2000)
高温、酸性	干含硫土壤	*Picrophilus oshimae/torridus* (A), 硫磺矿硫化叶菌(A), 热原体(A)	pH_{opt}0.7 (1.2mol/L H_2SO_4)	Schleper et al.(1995); She et al. (2001); Johnson (1998)

注：(A)，古生菌；(B)，细菌。

嗜热嗜酸微生物，例如硫化叶菌属（*Sulfolobus* spp）和热原体属（*Thermoplasma* spp）发现于硫质火山口和其他低pH的环境中，例如热煤废弃物中。这些热环境中只有原核生物存在，而没有真核生物的竞争。已知的最嗜热的真核生物是纤毛虫，但它们不能在超过52℃的地热泉水中生长[3]。尽管嗜热微生物有宽广的温度生长范围，但它们与正常微生物一样采用基本的基因组结构和编码习惯。因而全基因组序列分析就是探

索高温生存和生长的分子基础的最直接的方法。

本章主要介绍于2003年5月之前完成基因组测序的15种嗜热微生物及其基因组(表2)。这些完成的基因组测序项目以时间顺序来显示，并且包括菌株的最佳生长温度以及它们能量储存的主要方式。

表2 已测序的嗜热微生物基因组

种名	基因组大小	完成时间	T_{opt}	能量贮存方式
詹氏甲烷球菌(*Methanococcus jannaschii*)	1.66	1996(5)	80	甲烷产生
闪烁古球菌VC16(*Archaeoglobus fulgidus* VC16)	2.18	1997 (6)	83	硫酸根还原
热自养甲烷杆菌ΔH(*Methanobacterium thermoautotrophicum* ΔH)	1.75	1997 (16)	65	甲烷产生
霍氏火球菌OT3(*Pyrococcus horikoshii* OT3)	1.73	1998(18)	98	异养/S_0还原
敏捷气热菌(*Aeropyrum pernix* K1)	1.80	1999 (25)	95	异养
风产液菌(*Aquifex aeolicus*) (B)	1.80	1999 (9)	85	H_2氧化
海栖热袍菌(*Thermotoga maritima*)MSB8 (B)	1.86	1999 (8)	80	异养,S_0还原
深海火球菌(*Pyrococcus abyssi*)	1.75	2000 (19)	95	异养,S_0还原
激烈火球菌(*Pyrococcus furiosus*)	1.91	2000 (17)	100	异养,S_0还原
Sulfolobus tokodaii 7	2.69	2000 (13)	80	异养,S氧化
硫磺矿硫化叶菌(*Sulfolobus solfataricus* P2)	2.99	2001 (12)	85	异养,S氧化
嗜酸热原体(*Thermoplasma acidophilum* GSS1)	1.58	2001 (21)	60	异养,S氧化
嗜酸热原体(*Thermoplasma acidophilum*)	1.58	2001 (20)	63	异养,S氧化
Methanopyrum kandleri AV19	1.69	2001 (22)	103	甲烷产生
需氧热棒菌(*Pyrobaculum aerophilum* IM2)	2.22	2002 (24)	100	硝酸盐还原
Thermoanaerobacter tencongensis (B)	2.69	2002 (10)	75	发酵,S还原

注:T_{opt},最适生长温度;(B),细菌。

值得注意的是，随着具有历史意义的第一个微生物基因组，病原微生物流感嗜血菌(*Haemophilus influenzae*)在1995年的测序[4]，接下来完成的全基因组测序的是深海嗜压超嗜热微生物詹氏甲烷球菌(*Methanococcus jannaschii*)[5]。从那以后，每年都有几个嗜热微生物的基因组测序完成。

嗜热、极端嗜热和超嗜热微生物

嗜热微生物的定义是在较高温度条件下才能生长的微生物。这是不同于例如形成芽孢的耐热微生物，它们采用特殊的生存策略以至于它们能够在较高的温度条件下生存但却不能生长。下面的称谓用来区分嗜热微生物的不同适应能力：能够生长于50～75℃之间的微生物一般称之为嗜热微生物(thermophiles)，生长于75～90℃之间的微生物为极端嗜热微生物(extreme thermophiles)；生长于90℃以上的微生物一般称之为超嗜热微生物(hyperthermophiles)[6]。目前关于超嗜热微生物的描述是相当多样的，已知23个属中都有超嗜热微生物的成员[7]。有几个属(例如，火叶菌 *Pyrolobus*、热网菌 *Pyrodictium*、热棒菌 *Pyrobaculum*、超高温甲烷菌 *Methanopyrus*、火球菌 *Pyrococcus*、硫化

叶菌 *Sulfolobus* 和古球菌 *Archaeoglobus*）都是超嗜热微生物。虽然有几种细菌也被认为是超嗜热微生物，但最佳生长温度在 90℃ 或以上都是古生菌[6,7]。已完成基因组测序的超嗜热细菌有海栖热袍菌（*Thermotoga maritima*）[8]、风产液菌（*Aquifex aeolicus*）[9]和 *Thermoanaerobacter tencongensis*[10]。

由于它们的高生长温度，许多通常使用的遗传学工具和遗传分析方法是不能用来研究超嗜热微生物和许多极端嗜热微生物的。筛选方法，例如抗生素的抗性，是不能在高温下使用的，这是因为多数抗生素是热不稳定的或者分解抗生素的蛋白质是热敏感的。虽然有几种方法能够使用抵抗高温的凝胶例如 Gellan 胶或硅胶替代琼脂来制备固体培养基供极端嗜热微生物生长[11]，但在有些情况下，要么是无法形成菌落，要么是菌落形成太慢而不适合于遗传分析。在这种情况下，超嗜热微生物的基因组测序就成了唯一可行的遗传分析方法。但硫化叶菌例外，该属目前有两个全部基因组测序的种：硫磺矿硫化叶菌（*Sulfolobus solfataricus*）[12]和 *Sulfolobus tokodaii*[13]；这些生物好氧，较易培养，在 80℃ 容易形成菌落。可以用人工构建的重组表达载体和热稳定性的抗生素潮霉素（hydromycin）对它们进行筛选[14]。最小的古生菌细胞是专性共生的 *Nanoarcheum equitans*；由于它专性共生于嗜热微生物寄主 *Ignicoccus* spp 上而得名[15]。这个菌株的基因组小于 500kb，并且它的 16S 核糖体核糖核酸（SrRNA）序列建议它可能是新的门——微小古生菌门（*Nanoarcheota*）。

用比较基因组学研究嗜热微生物的热适应性要求

目前已有多个基因组序列可用于亲缘关系相近的嗜热微生物研究，包括三种嗜热甲烷菌：詹氏甲烷球菌[5]、热自养甲烷杆菌（*Methanobacterium thermoautotrophicum*）[16]和 *Methanopyrus kandleri*[17]；两种硫化叶菌：*S. tokodaii*、硫磺矿硫化叶菌 P1[13]；三种火球菌[18~20]和两种热原体菌[21,22]。比较基因组研究已经成为理解微生物在高温环境中生存和繁殖的适应性的非常重要的因素。表 2 也提及了最近完成的高温超嗜热微生物 *M. kandleri* 的测序[17]，该甲烷菌最高能在 110℃ 生长[23]。另外还有需氧热棒菌（*Pyrobaculum aerophilum*）；它是微好气，通过硝酸还原在 100℃ 下最佳生长[24]。最后，一个专性好气并且能够在最高温度 100℃ 生长的微生物，敏捷气热菌（*Aeropyrum pernix*），在 1999 完成了测序[25]。

因为它们的个体小，各种类型的嗜热微生物的细胞内成分在生长过程中暴露在周围环境中。为了使生化过程在高温下正常进行，所有的细胞内成分都必须有显著的内在热稳定性。在许多情况下，这是因为适应性提供了高温下的内在稳定性；例如，来自超嗜热微生物的蛋白质几乎毫无例外的有高的内在稳定性。然而，有些成分如 DNA、RNA 和生化路径中的许多中间产物的纯化形式在其来源生物的最佳生长温度是相当不稳定的。在有些情况下，中间产物（例如氨基甲酰磷酸）是极端不稳定的，此时某些适应性的酶结构便通过在酶的复合体内运输中间产物而使反应路径得以进行，从而避免了中间产物在细胞质中的降解[26]。当转化过程快速进行时，细胞内的条件和适应性策略还必须保证中间产物持久性和可替换性。为了使细胞得以生存，超嗜热微生物的细胞比体外的单个蛋白质和蛋白质核酸复合物有更强的热稳定，在有些情况下甚至有 20℃ 的差别。

下面部分涵盖了在超嗜热微生物中的细胞成分的适应稳定性和修复机制。

DNA 的修补和复制

超嗜热微生物的令人着迷的特征之一，是它们能够在远远超过它们的 DNA 和 RNA 的变性温度下进行生长。尽管嗜热性细菌常常因为基因组有较高的 G+C 含量从而升高了 DNA 的熔点温度，但是超嗜热微生物的基因组 DNA 的 G+C 含量却分布在比较宽的范围内，而且大多数超嗜热微生物的基因组的 G+C 含量相对较低。*M. kandleri* 则是个例外，它的 G+C 含量是 65%[23]。

对于这个奇特现象目前有两种解释。通常，DNA 或 RNA 的熔点温度的测量是在标准溶液例如标准柠檬酸盐溶液中进行的，而熔点温度在较高的离子强度和在某些有机溶质中存在的情况下将随之升高。在全部已研究的超嗜热微生物中，它们细胞质中都含有高浓度的相应溶质，例如，激烈火球菌的胞内谷氨酸钾浓度是 200～600mmol/L[18]，钾的反离子可以是谷氨酸离子，也可以是磷酸肌醇（di-myo-inositol-1, 1'-phosphate）离子。*Methanopyrum kandleri* 能够在高达 110℃下生长但却不能在低于 90℃下生长，它的细胞质中含有浓度高达 2.5M 的环式双磷酸甘油酸[17,23]。事实上，因为超嗜热微生物的细胞质的溶质浓度比它们外界环境高，所以可以把它们归于“非正式的嗜盐微生物 (closet halophiles)”。这些溶质在维持超嗜热微生物的酶和 DNA 双螺旋的稳定性上起非常关键的作用[27]。高盐在体外也能够防止 DNA 断裂；相对于有缺刻的或线状的 DNA 而言，共价闭合环状的 DNA 不易断裂[27]。

温度能够使 DNA 双螺旋发生变化，因此温度的变化本质上将改变复制子的超螺旋密度[27]。人们对几种超嗜热微生物中的 DNA 拓扑酶的活性进行了研究。许多嗜热微生物有编码逆促旋酶（reverse gyrase）的 *rgy* 基因的同源物（homolog），它存在于所有超嗜热的古生菌和细菌中[28~31]。最近，Forterre[31]使用比较基因组方法发现了超嗜热微生物中特有的基因直系同源群簇（cluster of orthologous group，COG）。这项研究仅仅发现了逆促旋酶，该酶在维持在中性或正超螺旋状态下的染色体和质粒的超螺旋密度上起重要作用[29]。这些酶有独特的双功能性，它含有解旋酶的区域和拓扑酶 I 的区域[30]。当逆促旋酶在 DNA 链上移动时，解旋酶在前面产生正超螺旋，拓扑酶在后面释放负超螺旋，结果在复制元中产生了净的正超螺旋[30]。超螺旋密度的调控机制一直不太清楚。*M. kandleri* 的逆促旋酶是非常独特的，它具有高活性并且由两个亚基组成[32]（其他逆促旋酶都含有两个功能域的单个蛋白质）。*M. kandleri* 的另一个拓扑酶 V 也表现出双功能酶活，它的重要功能是在高温时碱基切除修复中切开 DNA 骨架[33]。

Euryarcheota 的 DNA 结合蛋白中有一种是古生菌组蛋白（archaeal histone），它形成四聚物核小体维持 DNA 在高盐条件下的正超螺旋，在体外稳定和压缩 DNA。詹氏甲烷球菌的组蛋白很特殊，它们有 C 端的扩展序列，该序列并不干预核小体形成，反而大大增强了核小体的稳定性[35,36]。*Crenarcheota* 的 DNA 结合蛋白（包括新型染色质形成蛋白 Alba[37]），通过提高核蛋白复合体的熔点温度增加 DNA 稳定性，与真核生物的组蛋白类似，Alba 的结合特性可以通过乙酰化调节[38]。核蛋白复合物体一方面通过紧密结合保护 DNA，另一方面根据转录和翻译的需要释放 DNA，这样既可以阻止 DNA 在高温时水解断裂[39]，又可以保证快速而精确进行 DNA 双链的破裂修补[40]。有趣的

是，嗜酸热原体（*Thermoplasma acidophilum*）[21]和火山热原体（*Thermoplasma volcanium*）[22]的基因组序列，不含有类似 *Crenarcheota* 或*Euryarcheota* 的古生菌组蛋白，取而代之的是细菌基本 DNA 结合蛋白 HU，这可能是基因水平转移（lateral gene transfer, LGT）的结果，如是就产生了一个有趣的问题，热原体的温度范围是上升了，还是下调了呢？

古生菌基因组中 DNA 复制所需的大多数基因都已经发现[41]，在古生菌中普遍发现的复制成分是增殖性细胞核抗原（proliferating cell nuclear antigen, PCNA）家族，以及解旋酶的微小染色体维持蛋白族（minichromosome maintenance class），前者能在复制叉形成环状夹子，后者促使复制叉分开成链[42]。由于与真核生物复制叉成分相近，古生菌的复制过程与真核生物相似，超嗜热真细菌的 DNA 复制则采用众所周知的细菌复制成分，推动复制叉前行的 PCNA 蛋白在 *Euryarcheota* 中只有一个拷贝，但是，在 *Crenarcheota* 中却有三个相近的同构体（isolog）[42]。已经鉴定了许多与复制相关的 DNA 多聚酶，Cann 等[43]在 *Euryarcheota* 中发现了一类独特有复制活性的 DNA 多聚酶，它们由二聚体组成，与细菌或真核生物的多聚酶序列没有明显相似性，其他多聚酶在已测序的 *Euryarcheota* 基因组中具有保守性[44]，另外的 B 类 DNA 多聚酶也在 *Euryarcheota* 中发现，它们在滞后链的合成和 DNA 修补中发挥作用。在 *Crenarcheota* 中，B 类多聚酶在修补和复制两方面都起作用[41]，经常以多个并系同源基因（paralogous gene）形式存在。

与酵母的连接酶一样，古生菌连接酶在序列上与典型依靠 ATP 的真核生物酶[45]相似；而嗜热细菌的连接酶是典型依靠 NAD 的酶，只不过有内在高度稳定性[46]。

超嗜热微生物的超常 DNA 断裂修补能力，表明它们在正常生长状态下有活跃的重组系统[40]。对硫磺矿硫化叶菌中的霍利迪连接体（Holliday junction, Hjc）的分解酶研究[47]显示，超嗜热微生物的同源重组系统，对它们的双链断裂修补能力非常关键。Hjc 酶是一个依赖分支的核酸酶，它能酶解霍利迪连接体并能把重组链分开。对热稳定 DNA 修饰酶的纯化和特性研究，使得用中温酶难以或根本无法进行的那些生化研究能够开展。

热激或冷激

在地热或热泉环境中生活的嗜热微生物，经常暴露在热的循环中，既有高于又有低于最佳生长温度的时候，环境条件的改变使它们不断处于热激和冷激（heat and cold shock）的交替变化中，因而，在嗜热微生物基因组中发现为数不多的热激或冷激基因令人感到惊奇。超嗜热微生物激烈火球菌中的超氧化物还原酶（superoxide reductase）是冷适应蛋白中的一种，它在温度低于最低生长温度 80℃ 时仍有活性[49]，这可能与这些氧敏感细胞从火山口中喷出而进入冷的、有氧的海水中时，对抗氧损伤的一种适应性反应。

嗜热细菌有 groE/groEL 和 HSP70 分子伴侣的全套分子，有趣的是，在 *Euryarcheota* 基因组中也发现了几个这样的基因，例如，甲烷菌和嗜盐菌（综述见参考文献［50]）。然而，超嗜热微生物似乎不编码这些热激系统，取而代之的是依靠 ATP 桶状 HSP60 分子伴侣（常被称为热体，themosome）和只在超过最佳温度时才产生的一个

小热激蛋白[51]。该分子伴侣类似脊椎动物眼睛晶状体（eye-len）蛋白中间区域的 α-晶状体球蛋白同源体，它已从詹氏甲烷球菌中得到结晶[52]，而激烈火球菌的该分子伴侣在大肠杆菌中表达后，寄主细胞能在 50℃ 长时间成活[53]。

蛋白质的热稳定性

发现生活在温度接近或高于 100℃ 的超嗜热微生物，已引起对蛋白质热稳定分子机制的许多研究，是否有适用所有超稳定性蛋白的通用法则，已引发了热烈的讨论，如果有，是否这些法则可用来设计更加热稳定的有用酶？已经发现普遍适用于热稳定蛋白的几个特性，例如，由于它们大多有较小的环和较短的 N 端和 C 端，所以蛋白质内部空隙较少，蛋白结构也更紧凑。在很多情况下，与中温酶相比，热稳定酶多以寡聚体的形式存在，当寡聚体被定点突变破坏后，抗热性下降，寡聚体越大，越容易在亚基之间运输底物，这可能是超嗜热微生物普遍使用的策略[54]，而热稳定蛋白的亚基比中温酶的亚基小[5]（综述见参考文献 [55] 和 [56]）。

詹氏甲烷球菌是从深海热泉排出口中分离的嗜压微生物，也是第一个完成测序的超嗜热微生物[5]，它只有在 50mPa 压力下才能达到最大生长速率和温度上限[57]（表 2），它的基因组信息为与中温微生物进行比较蛋白质组研究提供了高温标准，对高压和高温下的蛋白质适应性研究显示，低于变性压力的中等压力（≤100mPa），能极大地增强蛋白质对热的抗性[58,59]，该效应在温度很高时尤为明显，这种现象对嗜热微生物蛋白质在极端环境下的适应性起重要作用。

离子对的相互作用对超稳定蛋白质的贡献已经争论了好几年，在理论上，静电相互作用对蛋白质结构的内部成键十分有利，这是因为对于范德华力它们有较长的作用范围（最多到 4Å），而对在接近或超过 100℃ 时水结构变弱（这将使疏水效应变弱）也不敏感。从相近类别中温和嗜热微生物蛋白质组的比较研究得知，较高生长温度与较高比例带电氨基酸残基呈正相关[60,61]，实验和理论研究都发现离子对的相互作用大多起稳定作用[62~64]。

越来越多超嗜热微生物的蛋白质晶体结构已经阐明，并与相应热不稳定蛋白质进行了比较[65,66]，在有些情况下，在不太稳定的蛋白质中放置离子可以提高其稳定性[67,68]，几种蛋白质结构的比较研究已确证，超嗜热微生物同源物离子对的相互作用非常丰富的，基因组比较研究也支持这个结论。

热原体属中两个基因组测序的完成很有价值，因为这些菌株的生长温度相对较低（60℃ 左右），因此它们的蛋白质组提供了与高温菌株相比的低温标准[21,22]。火山热原体的氨基酸组成分析印证了两个比较研究的结论[60,61]，普遍的结论是随着生长温度升高，带电氨基酸残基（如 Glu、Asp、Arg 和 Lys）增加，而极性残基（如 Ser、Gln、Tyr 和 Asn）则相应减少。虽然，带电氨基酸残基明显涉及离子对的形成，但也有人认为超嗜热微生物蛋白质在细胞内的溶解性，可依赖于高盐条件下增加的表面电荷，从而导致蛋白质组总电荷的增加[60]，这就是盐杆菌菌株 NRC1 的蛋白质组富含带电氨基酸残基的原因（见第 21 章）。

结构基因组方法对解决这个问题有极大的贡献，在理论上有了推导的蛋白质组就意味着可以通过重组表达技术获得所有蛋白质。但实际上，这仅仅适用于那些能够在适当

寄主中表达、正确折叠和进行适当翻译后修饰的蛋白质，而膜蛋白通常不容易获得，除非它们能在亲缘关系相近的寄主系统中表达。

尽管有这些困难，海栖热袍菌[69]、詹氏甲烷球菌[70~74]和激烈火球菌[74]的高通量表达和晶体结构分析项目正在进行，当足够数量超稳定蛋白质的晶体结构解析后，通过与中温蛋白结构进行比较，就可以回答热稳定蛋白质的过量电荷是被掩埋，还是暴露在溶剂中。海栖热袍菌和激烈火球菌的许多蛋白质结构已被解析，因为这些微生物能高密度生长，因此，可以用常规方法从无细胞抽提物中纯化高表达量的天然蛋白质，通过这种方法，激烈火球菌的 100 多种天然蛋白质已经纯化和分析，这种方法的主要优点是用现成的附加基团获得对金属蛋白或糖蛋白的亲和性。超稳定蛋白质的正确折叠需要暴露在相似的溶剂浓度和温度中[74~76]，因此，许多超稳定蛋白质的重组表达需要一些新方案和新手段。

生物合成和代谢系统

嗜热微生物有多种自养和异养的能量贮存方式；一般来说，自养微生物有完全的生物合成核酸、氨基酸和维生素的能力。在异养嗜热微生物中，生物合成能力各不相同，许多微生物在生长中需要维生素和氨基酸[11]。因此这一节阐述能量贮存和生物合成能力之间的关系。

生物合成

对于接近生存上限的生物系统，它们的细胞质膜含有称为古生菌脂肪（archaeal lipid）的物质，古生菌膜脂肪很特殊，它们的萜类或植烷链是以醚键代替酯键，连接脂肪族链和甘油衍生物的极性基团。以醚键连接的双醚和四醚脂肪在超嗜热微生物的脂肪中尤为普遍[6]，含有高双醚和四醚脂肪的膜是单分子层，而不是熟知的双分子层，它们对极端温度和氧化/还原有无与伦比的稳定性，并能在极端 pH 和温度中保持质子的不通透性[11]。

脂肪的生物合成十分有趣，因为古生菌的醚脂需要类异戊二烯生物合成途径，其中 3 羟基-3-甲基戊二酰辅酶 A 还原酶（3-hydroxy-3-methylglutaryl-CoA reducase，HMG-CoA reductase，E.C. 1.1.1.34）是关键酶。在真核生物中，HMG-CoA 还原酶是胆固醇生物合成速度的限制因子，也是降低胆固醇许多药物的靶标；在细菌中，这种酶经常是产生萜类化合物抗生素非必要途径的部分[78]。有趣的是，闪烁古球菌（*Archaeoglobus fulgidus*）和热原体的基因组含有细菌的 HMG-CoA 还原酶，而在其他古生菌中则是真核生物/古生菌的 HMG-CoA 还原酶[74]，看来细菌的基因替代了古生菌的基因，这是基因丢失而影响重要功能的一个例子[80]。

异养嗜热微生物的基因有失也有得，例如，火球菌的氨基酸生物合成途径和对碳水化合物（如纤维二糖、麦芽糖、海藻糖、海带多糖和几丁质）吸收与合成代谢有显著的多样性。激烈火球菌基因组（1.95Mbp）明显比深海火球菌（*Pyrococcus abyssi*）（1.75Mb）和霍氏火球菌（*Pyrococcus horikoshii*）（1.73Mb）的基因组大；激烈火球菌和深海火球菌的基因组都编码操纵子 *trp*、*aro*、*arg* 和 *lie*/*val*，但霍氏火球菌基因组

却没有这些功能；深海火球菌和霍氏火球菌基因组都缺少组氨酸和维生素 B12 的合成途径、部分三羧基循环途径以及对淀粉、麦芽糖、海带多糖和纤维二糖的发酵（吸收和降解）途径；霍氏火球菌是火球菌中最小的基因组，它是 Val、Leu、Ile 的营养缺陷型，可能也是其他芳香类氨基酸的营养缺陷型，因为它缺少通常的芳香化合物合成酶；激烈火球菌能合成全部核苷酸和维生素 B12 和 B6，能在缺少这些营养物的合成培养基中生长[82]。这些特性反映了这些具有生物合成能力的基因的存在。Ettema 等[83]认为，如同细菌的基因组一样，超嗜热微生物的基因组中也有整块的染色体片段的插入和删除。

这可能有如下的暗示：假设火球菌的深海菌株所生存的环境很少含有糖类，而激烈火球菌的海岸生存环境通常是高糖低多肽的情况。由于激烈火球菌能够有效地利用许多碳水化合物，它在缺少氨基酸时可利用糖的底物的生长，这可能导致了氨基酸生物合成途径的引入；而这些途径在霍氏火球菌和深海火球菌中是没有的。

这三个非常相似的基因组产生了一个有趣的问题：是深海火球菌沉降到深海而导致了它们基因组的减小和代谢多样性的萎缩呢？还是激烈火球菌从海底上升而获得新的代谢能力呢？目前的证据支持后一种假说。

虽然基因组可塑性的机制没有被发现，在其他火球菌中缺乏激烈火球菌的麦芽糖区域，说明这是新近引入的基因组区域[84]。这个含有麦芽糖操纵子的 19kb 片段在基因顺序和整体序列上都类似于大肠杆菌的 *mal* 操纵子，并且这个片段的侧翼有 2 个插入序列（insertion sequence, IS）（激烈火球菌基因组中共有 23 个 IS)。与相近的古生菌 *Thermococcus litoralis* 的麦芽糖操纵子区域比较，发现两个菌株的整个区域是非常相似的。激烈火球菌和 *T. litoralis* 的这个区域在进化上形成分枝后，仅仅形成了 196 个突变。对转座子的末端正向重复序列的分析显示这个区域是作为复合转座子进行转移的[84]。

火球菌中 IS 元件的分布是十分多样的。激烈火球菌至少有 29 个全长度的 IS 元件[18]。其中有一个 IS 元件由两个完整 IS 元件串联组成。虽然 *Thermococcus litoralis* 比火球菌的生长温度低，基因转移还是有可能的，因为它们的栖息地是重叠的，并且 *Thermococcus litoralis* 与激烈火球菌分离自同一地点（意大利的海岸线)[49]。

中心代谢

超嗜热微生物的中间代谢的一个不寻常的特征是经常在反应中以 ADP 替代 ATP。例如，火球菌中 EMP 途径好像缺失和功能不全，因为找不到依赖 ATP 的磷酸果糖激酶和葡萄糖激酶。但灵敏的生化分析最终被发现了这些途径中的依赖 ADP 的酶[85]。这些酶已经被纯化、克隆和表达[86,87]。当基因组序列被重新分析时，在中温和嗜热性的甲烷菌中都发现了依赖 ADP 的激酶[88,89]。

这种叙述可能使读者认为超嗜热微生物因为热稳定性的原因而在 EMP 途径中限制性地使用 ADP 来替代 ATP。但这并不一定正确；好气性超嗜热微生物敏捷气热菌就有正常的 B 族依赖 ATP 的磷酸果糖激酶[90]。无论如何，超嗜热微生物如何在正常代谢流程产生过量的 ADP 的问题以及激烈火球菌[91]、詹氏甲烷球菌和闪烁古球菌[92]中 ADP 形成的乙酰辅酶 A 合成酶的描述可能说明在超嗜热微生物中有产生 ADP 的新步骤，这就有可能发现一些具有不寻常代谢功能的新基因。

氢代谢是超嗜热微生物的代谢上的一个被深入研究的特点和重要方面。嗜热的产氢古生菌和细菌编码的氢化酶有的在膜上，有的在细胞质中（综述见参考文献 [93]）。膜固定的氢化酶含有 10 到 14 个亚基，它们能将胞内的质子转换成为氢气并释放到胞外[94]。在激烈火球菌中，基因组含有两个推定的操纵子，它们非常类似于细菌（如深红红螺菌 *Rhodospirillum rubrum*）中的氢化酶操纵子（见第 12 章)。但是，激烈火球菌的氢化酶复合体的大亚基中有一个独特的结合镍铁簇的模体（motif)。对于膜上氢化酶，可溶性的氢化酶以及硫化物脱氢化酶的活性的动力学研究显示，激烈火球菌中进行能量运输、转移、再生和调控的细胞装置把具有还原力的三类分子（铁氧化还原蛋白、烟酰胺腺嘌呤二核苷酸磷酸，H_2）联系在了一起[93,95]。

膜和运输系统

嗜热微生物在选择性渗透运输系统的操作中面临两个挑战。首先是膜完整性的控制，如同在生物合成的那节中所说的那样，这是通过跨膜的脂类来完成的[96]。受热条件下的二醚和四醚脂肪的含量被上调，从而影响到质子的渗透性；这种机制在细菌中没有被发现[96]。这可能是细菌的耐高温能力比古生菌弱的原因。两个硫化叶菌[12,13]的基因组序列的公开使人们将这些嗜热嗜酸菌的信号序列分成四类：分泌信号、双精氨酸信号肽、脂蛋白信号肽和 IV 型类菌毛信号肽[97]。硫磺矿硫化叶菌的蛋白质组中 4.2% 的蛋白质是信号肽，比詹氏甲烷球菌的蛋白质组（2%）高得多。

分泌蛋白的目的地由信号肽控制，大多依赖于（ATP-binding cassette，ABC）结合蛋白[98]。这些重要的转运蛋白可分为两类：一类只含有单个的 ATP 酶，它主要是负责单糖的吸收；另一个含有两个 ATP 酶，也可吸收二糖和低聚糖[99]。类似的系统已经在 Euryarcheota 中有所报道，例如 *T. litoralis* 和激烈火球菌中的麦芽糖吸收系统[84]。古生菌中的许多吸收系统是不依赖 ATP 的细胞间质类型，它们分布广泛，但是却没有在真核生物中发现。此类运输所需的能量不是来自 ATP，而是来自电化学离子梯度[100]。

相对于许多中温菌系统而言，超嗜热古生菌的吸收系统的一个显著特征是具有极高的底物亲和能力[98]。高温对糖类有强破坏能力，特别是对单糖，因此吸收系统必须极其有效地捕获全部的单糖。糖类和肽的寡聚物的吸收也是极其有效的，并且主要是被 ABC 转运子或不依赖 ATP 的细胞间质系统来介导。后者由钠或质子梯度来驱动。有些分离株仅靠吸收寡肽来生长；例如，Gonzalez 等描述了一种深海异养菌 *Thermococcus peptonophilus*；它在所有的热球菌（*Thermococcus*）中生长得最快，但不能在单个的氨基酸中生长[101]。

基因组的可塑性和种系关系

水平基因转移

嗜热的细菌和古生菌经常栖息在同样的生态环境中，来自基因组分析的大量证据显示在它们之间普遍存在着水平基因转移。例如，海栖热袍菌的基因组估计有 25% 基因与超嗜热古生菌同源，特别是与火球菌属[8]。在海栖热袍菌中具有细菌特征的基因占

优势，但它的基因组中也明显存在着来自古生菌的基因[102]。

超嗜热古生菌有极端的密码子偏好性，例如超嗜热古生菌只使用 AGG 和 AGA 编码 Arg，而且不用 UAU 来编码 Ile[103]。因此，如果要在细菌中表达超嗜热古生菌的蛋白质，就需要用 pSJS 这样的质粒在细菌中同时异源表达稀少编码的转运 RNA（tRNA），以调整密码子使用率；用这种策略已经在大肠杆菌中成功地过量表达超嗜热微生物中的重组蛋白[53]（S. Sandler，个人通讯）。极端嗜热细菌 *Carboxydothermus hydrogenoformans* 编码一氧化碳脱氢酶的基因已经被克隆了，并且发现它是古生菌中的一氧化碳脱氢酶的主要同源物，它有典型的古生菌的密码子偏好性[104]。

这就提出了一个问题：LGT 是否广泛存在于共同生活在同一小生境的微生物中？如果是，那么尝试使用个别基因来构建种系关系树将是徒劳，因为谁也不能确定该基因是从祖先继承来的。解决这个问题的一个办法就是利用日益增多的基因组序列：抽取和比较基因组来构建种系关系树[105]。使用 SHOT（短直系同源和基因顺序树构建工具，short ortholog and gene order tree construction tool，SHOT；网上工具可构建这样的树[106]；http://www.bork.embl-heidelberg.de/SHOT）。

图 1 是利用现有的嗜热微生物的基因组来构建的；盐杆菌 NRC1 的基因组在这里作为一个外部对照。总的来说，由于嗜热微生物有相似大小的基因组，所以用这种方法来研究它们的种系关系是实际可行的。虽然基因组的内容可能会因为 LGT 而混杂有外来基因，但这些树至少提供了研究菌株的相似性的基础。

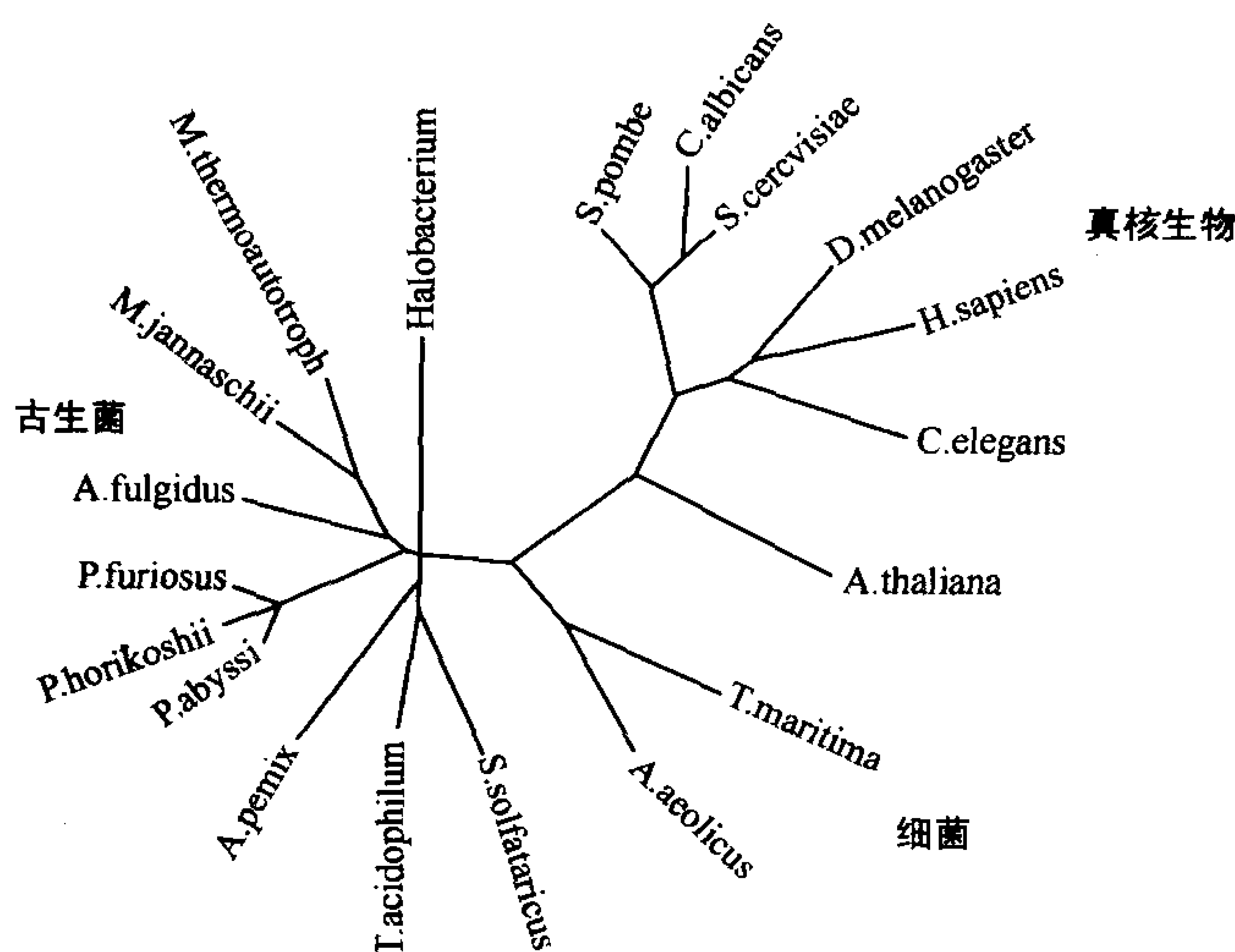


图 1　与中温生物对照，采用 SHOT[107]方法对嗜热微生物的全基因组的分析结果。

在图 1 中，嗜酸的热原体和硫化叶菌是成簇的，如同 Methanogen-Archaeoglobus 组一样。好气的 Crenarcheote 同其他的好气古生菌形成较深的分支。这些菌株的主要生理特点也支持这种关系。

这个研究的缺陷是盐杆菌 NRC1 的位置，它占据非常深的分支，在古生菌的“真细菌”末端；这与 16 SrRNA 和 r-蛋白的分类相矛盾，在那里，嗜盐古生菌位于 eur-

yarchaeal 树的顶端。16 SrRNA 的分类显示盐杆菌 NRC1 与闪烁古球菌和詹氏甲烷球菌的亲缘关系最相近。但是，盐杆菌 NRC1 的许多基因与革兰氏阳性菌枯草芽孢杆菌和耐辐射异常球菌（*Deinococcus radiodurans*）又最相似，说明 NRC1 通过 LGT 获得了相当数量的基因（见第 21 章）。

使用全基因组分析也有缺陷，这是因为基因次序的保守性不高，特别是在自养超嗜热微生物中，例如风产液菌、詹氏甲烷球菌和 *M. kandleri*[5,9,17]。

甲烷产生菌 *M. kandleri* 在古生菌 16S rRNA 分类树中的放置较为反常，它位于古生菌分支的根部，其余的甲烷产生菌位于 Euryarcheota 的顶部。*Methanopyrum kandleri* 在压力存在下最高可在 110℃ 生长，这是目前已测序的嗜热古生菌中的最高记录。另外该菌有甲烷菌中最高的基因组 G+C 含量。有趣的是，最近的基于核糖体蛋白的分析强烈认为该菌属于甲烷菌/热球菌的分支[17]。

人们通常认为，编码核糖体蛋白的基因一般不会在不相关的菌株间转移。由于其 rRNA 的 G+C 含量极高（75%），*M. kandleri* 在 16S rRNA 的分类树中的放置看来是人为的。基因组序列显示 *M. kandleri* 的甲烷产生基因在氨基酸序列上与其他甲烷菌的相似，而且基因次序也有保守性。这些都说明 *M. kandleri* 是典型的甲烷产生菌，并且所有的甲烷产生菌都属于一个分类群。

极高的 G+C 含量很可能是 *M. kandleri* 对高温生长的一种适应。对 *T. tencongensis* 的基因组分析显示基因组的 G+C 含量与 rRNA 和 tRNA 的 G+C 含量无关[10]。总的来说，生长温度与 RNA 的 G+C 含量（在嗜热微生物中可达 65%～70%）有十分密切的相关。例如，深海火球菌的基因组的 G+C 含量是 42%，而 rRNA 的 G+C 含量是 68%。正如同 Galtier 等所指出的那样[105]，这会人为地导致超嗜热微生物的以 rRNA 为基础的分类错误，因此对超嗜热微生物的共同祖先的认定要持保守态度。

超嗜热细菌海栖热袍菌的基因组中有很多可读框架（24%）与古生菌中的同源[8]，这说明在细菌和古生菌之间水平基因转移是广泛存在的。在这些分享的基因中，几乎有一半与火球菌的同源体有极大的相似性。栖热袍菌和火球菌共存于相同的海洋热水的小生境中，并且它们的生理特点是非常相似的（例如，它们都是异养的，能降解硫，能在超过 90℃ 时生长）。这些域内（例如热球菌－火球菌）和域间（例如栖热袍菌－火球菌）的水平基因转移说明生长在相同的条件下的嗜热微生物能够互相交换基因。

M. kandleri 的紧凑的基因组上几乎没有水平基因转移的痕迹[17]。这就进一步证实了共同的栖息地是诱导水平基因转移的条件。在生理上，*M. kandleri* 是位于生长温度范围顶部的“孤独骑士”，它在活跃生长期间没有太多的机会与其他的活细胞相接触。

M. kandleri 的另一个非同寻常的生理特性，是在它的细胞质中有极高浓度的磷酸和环式双磷酸甘油酸（cyclic diphosphoglycerate）。一个进入新细胞的蛋白质是否有功能取决于细胞质的组成成分。*M. kandleri* 的蛋白质只有在摩尔浓度的磷酸盐中才能发挥最佳功能；低磷酸盐浓度下，酶的稳定性、活性和亚基的组装都会受到负面影响；这称为液向性条件（lyotropic condition）[107～109]。这很可能也是由 LGT 整合新基因的一个巨大障碍。

插入元件和内含子

许多嗜热微生物的基因组总体上都缺乏保守的基因顺序，这就提出了染色体快速重

排的机制问题。IS元件是单向重组的明显来源，此外它同重复序列一样，可以作为局部同源序列来通过交互重组的方式促进插入、删除以及倒置等事件的发生。IS元件已经在嗜热微生物的基因组中被广泛报道，并且也实验证据显示它们仍然十分活跃。例如在硫磺矿硫化叶菌的基因组中富含几个IS元件家族（总共201），它在 *ura* 基因中的插入频率和插入位点能够通过氟乳清酸的阳性选择而被观测到[110]。在培养嗜热微生物时，有人发现插入元件的存在与基因组重排有相关性（R.Garret，个人通讯，2002）。

已知的IS元件分成两类主要类型，自身IS元件（autonomous）和非自身的类似于微型倒置重复单元的元件（non-autonomous miniature inverted repeat element-like element)[111]。后者利用完全IS元件的转运酶进行转移和并且在“乘客”序列的侧翼有正向重复。此外，DNA的病毒，质粒和远缘基因组的DNA片段也能够通过整合酶机制而整合，例如硫化叶菌病毒SSV1[112]、敏捷气热菌和霍氏火球菌的染色体上的类似于整合元的片段；这些类整合元片段位于DNA插入片段（例如质粒pXQ1）的两侧。

三个含有内含子的编码蛋白质的基因重要发现已经被报道了[113]。包括一个与真核生物着丝点结合因子5同源的蛋白和一个在敏捷气热菌、硫磺矿硫化叶菌和 *S. tokodaii* 中都存在的小核仁核糖核蛋白的亚基。这些内含子在RNA水平上通过典型的古生菌的螺旋-凸出-螺旋（helix-bulge-helix）机制被切除；这说明内含子有可能起源于古生菌。

未来趋势

超嗜热微生物的蛋白质组使结构基因组的研究成为可能。关于詹氏甲烷球菌、海栖热袍菌和激烈火球菌的高通量重组基因表达的研究正在进行。热原体和高温厌氧杆菌（*Thermoanaerobacter* spp.）等低温古生菌与超嗜热微生物的结构基因组的比较研究，将有利于阐明蛋白质热稳定性的机制。

在致死温度下引起细胞死亡的主要原因似乎是膜的完整性的丢失。在100℃以上保持膜完整性的研究在生化上已经逐渐清楚。古生菌脂肪中的环戊烷的数目随着温度的升高而增加[114]。通过这种环温度补偿（ring temperature compensation）机制，嗜热嗜酸古生菌的质膜能够维持质子的不渗透性和刚性的结构，最终在广泛的生长温度范围内维持细胞外小环境（pH值低于2.5）和细胞内部分（pH值6.5）之间有稳定和陡峭的质子梯度。这种机制的膜生物合成途径还不清楚；通过研究超嗜热微生物在温度升高时哪些基因会增强表达或许可以解决这个问题。

这就要用到以微阵列为基础的研究。微阵列已经被用来研究激烈火球菌[115,116]；这对研究细胞的整体应激反应是非常重要的。在激烈火球菌中调控热激反应的阻抑物的发现[112]（图2）和正在进行的对类似于细菌的Lrp调节子的研究[118]，将会使我们很快获得基因调控方面的信息。此外，通过极端嗜热菌（例如硫磺矿硫化叶菌和热自养甲烷杆菌）的研究，DNA复制叉细节的许多方面已经浮现，但复制起点功能的细节一直不清楚。这个问题的解决将为古生菌的细胞周期的调控机制提供有价值的信息。

最近发现中温古生菌使用终止密码子UAG和特殊的UAG-tRNA编码第22种氨基酸，吡咯赖氨酸（pyrrolysine)[119,120]。虽然还没有证据显示它在嗜热古生菌中的存在，

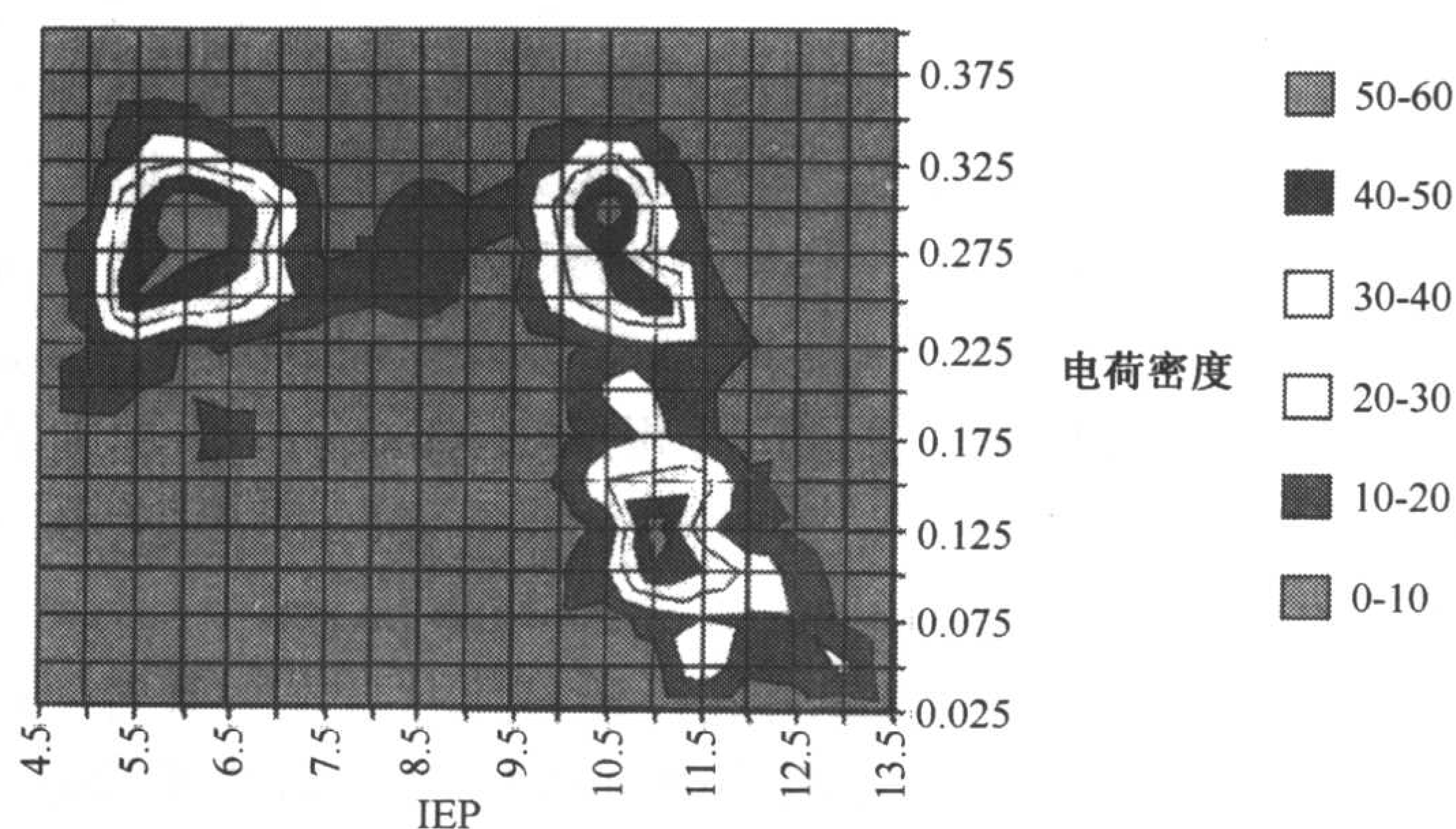


图 2　超嗜热微生物激烈火球菌的蛋白质组的电荷分布图。图密度，每块阴影面积中所含的蛋白质的数目。IEP，等电点。

但对已知的基因组重新研究可能会发现更多的 UAG 密码子与吡咯赖氨酸的关系。

对嗜热微生物的研究既有基础意义又有应用潜力。例如，这些细胞的能量生物转化与将来的非化石能源的生产相关，因为许多超嗜热微生物能够产生氢气或甲烷作为最终气态产物。对模式生物例如超嗜热微生物和嗜盐微生物的生长生理学和分子生物学的进一步研究，将有利于评估它们在生产气态能源和在生物技术领域中生产极端热稳定酶的潜力。

致谢

本章来自 Center for Maine Biotechnology 的第 04-612 号报告。作者感谢 NSF（项目编号 MCB02383387）和 Knut and Alice Wallenberg 基金会的支持。

（喻晓辉，刘　超　译）

参考文献

1. Madigan MT, Oren A. Thermophilic and halophilic extremophiles. Curr Opin Microbiol 1999; 2:265–269.
2. Blochl E, Rachel R, Burggraf S, Hafenbradl D, Jannasch HW, Stetter KO. *Pyrolobus fumarii*, gen and sp nov, represents a novel group of archaea, extending the upper temperature limit for life to 113°C. Extremophiles 1997; 1:14–21.
3. Baumgartner M, Stetter KO, Foissner W. Morphological, small subunit rRNA, and physiological characterization of *Trimyema minutum* (Kahl, 1931), an anaerobic ciliate from submarine hydrothermal vents growing from 28C to 52C. J Eukaryot Microbiol 2002; 49:227–238.
4. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
5. Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon,

Methanococcus jannaschii. Science 1996; 273:1058–1073.

6. Stetter KO. Extremophiles and their adaptation to hot environments. FEBS Lett 1999; 452:22–25.
7. Huber R, Huber H, Stetter KO. Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. FEMS Microbiol Rev 2000; 24:615–623.
8. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 1999; 399:323–329.
9. Deckert G, Warren PV, Gaasterland T, et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 1998; 392:353–358.
10. Bao Q, Tian Y, Li W, et al. A complete sequence of the *Thermoanaerobacter tengcongensis* genome. Genome Res 2002; 12:689–700.
11. Robb FT, DasSarma S, Place AR, Sowers KR, Schreier HJ, Fleischmann EM. (eds) Archaea: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1995.
12. She Q, Singh RK, Confalonieri F, et al. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci USA 2001; 98:7835–7840.
13. Kawarabayasi Y, Hino Y, Horikawa H, et al. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res 2001; 8:123–140.
14. Cannio R, Contursi P, Rossi M, Bartolucci S. Thermoadaptation of a mesophilic hygromycin B phosphotransferase by directed evolution in hyperthermophilic Archaea: selection of a stable genetic marker for DNA transfer into *Sulfolobus solfataricus*. Extremophiles 2001; 5: 153–159.
15. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 2002; 417:63–67.
16. Smith DR, Doucette-Stamm LA, Deloughery C, et al. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 1997; 179:7135–7155.
17. Slesarev AI, Mezhevaya KV, Makarova KS, et al. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci USA 2002; 99:4644–4649.
18. Robb FT, Maeder DL, Brown JR, et al. Genomic sequence of the hyperthermophile, *Pyrococcus furiosus:* implications for physiology and enzymology. Methods Enzymol 2001; 330: 134–157.
19. Kawarabayasi Y, Sawada M, Horikawa H, et al. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 1998; 5:147–155.
20. Cohen GN, Barbe V, Flament D, et al. An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol Microbiol 2003; 47:1495–1512.
21. Ruepp A, Graml W, Santos-Martinez ML, et al. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature 2000; 407:508–513.
22. Kawashima T, Amano N, Koike H, et al. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc Natl Acad Sci USA 2000; 97:14,257–14,262.
23. Burggraf S, Stetter KO, Rouviere P, Woese CR. *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. Syst Appl Microbiol 1991; 14:346–351.
24. Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH. Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc Natl Acad Sci USA 2002; 99:984–989.
25. Kawarabayasi Y, Hino Y, Horikawa H, et al. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 1999; 6:83–101, 145–152.

26. Massant J, Verstreken P, Durbecq V, et al. Metabolic channeling of carbamoyl phosphate, a thermolabile intermediate: evidence for physical interaction between carbamate kinase-like carbamoyl-phosphate synthetase and ornithine carbamoyltransferase from the hyperthermophile *Pyrococcus furiosus.* J Biol Chem 2002; 277:18,517–18,522.
27. Marguet E, Forterre P. Stability and manipulation of DNA at extreme temperatures. Methods Enzymol 2001; 334:205–215.
28. Confalonieri F, Elie C, Nadal M, de La Tour C, Forterre P, Duguet M. Reverse gyrase: a helicase-like domain and a type I topoisomerase in the same polypeptide. Proc Natl Acad Sci USA 1993; 90:4753–4757.
29. Bouthier de la Tour C, Portemer C, Kaltoum H, Duguet M. Reverse gyrase from the hyperthermophilic bacterium *Thermotoga maritima:* properties and gene structure. J Bacteriol 1998;180: 274–281.
30. Borges KM, Bergerat A, Bogert AM, DiRuggiero J, Forterre P, Robb FT. Characterization of the reverse gyrase from the hyperthermophilic archaeon *Pyrococcus furiosus.* J Bacteriol 1997; 179:1721–1726.
31. Forterre P. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet 2002; 18:236–237.
32. Kozyavkin SA, Krah R, Gellert M, Stetter KO, Lake JA, Slesarev AI. A reverse gyrase with an unusual structure. A type I DNA topoisomerase from the hyperthermophile *Methanopyrus kandleri* is a two-subunit protein. J Biol Chem 1994; 269:11,081–11,089.
33. Belova GI, Prasad R, Kozyavkin SA, Lake JA, Wilson SH, Slesarev AI. A type IB topoisomerase with DNA repair activities. Proc Natl Acad Sci USA 2001; 98:6015–6020.
34. Marc F, Sandman K, Lurz R, Reeve JN. Archaeal histone tetramerization determines DNA affinity and the direction of DNA supercoiling. J Biol Chem 2002; 277:30,879–30,886.
35. Sandman K, Bailey KA, Pereira SL, Soares D, Li WT, Reeve JN. Archaeal histones and nucleosomes. Methods Enzymol 2001; 334:116–129.
36. Li WT, Sandman K, Pereira SL, Reeve JN. MJ1647, an open reading frame in the genome of the hyperthermophile *Methanococcus jannaschii,* encodes a very thermostable archaeal histone with a C-terminal extension. Extremophiles 2000; 2:115–122.
37. Bell SD, Botting CH, Wardleworth BN, Jackson SP, White MF. The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation. Science 2002; 296:148–151.
38. Wardleworth BN, Russell RJ, Bell SD, Taylor GL, White MF. Structure of Alba: an archaeal chromatin protein modulated by acetylation. EMBO J 2002; 21:4654–4662.
39. Peak MJ, Robb FT, Peak JG. Extreme resistance to thermally induced DNA backbone breaks in the hyperthermophilic archaeon *Pyrococcus furiosus.* J Bacteriol 1995; 177:6316–6318.
40. DiRuggiero J, Santangelo N, Nackerdien Z, Ravel J, Robb FT. Repair of extensive ionizing-radiation DNA damage at 95°C in the hyperthermophilic archaeon *Pyrococcus furiosus.* J Bacteriol 1997; 179:4643–4645.
41. Bohlke K, Pisani FM, Rossi M, Antranikian G. Archaeal DNA replication: spotlight on a rapidly moving field. Extremophiles 2002; 6:1–14.
42. Daimon K, Kawarabayasi Y, Kikuchi H, Sako Y, Ishino Y. Three proliferating cell nuclear antigen-like proteins found in the hyperthermophilic archaeon *Aeropyrum pernix:* interactions with the two DNA polymerases. J Bacteriol 2002; 184:687–689.
43. Cann IK, Komori K, Toh H, Kanai S, Ishino Y. A heterodimeric DNA polymerase: evidence that members of Euryarchaeota possess a distinct DNA polymerase. Proc Natl Acad Sci USA 1998; 95:14,250–14,255.
44. Gueguen Y, Rolland JL, Lecompte O, et al. Characterization of two DNA polymerases from

the hyperthermophilic euryarchaeon *Pyrococcus abyssi*. Eur J Biochem 2001; 268:5961–5969.
45. Gunther S, Montes M, de DA, del VM, Atencia EA, Sillero A. Thermostable *Pyrococcus furiosus* DNA ligase catalyzes the synthesis of (di)nucleoside polyphosphates. Extremophiles 2002;6: 45–50.
46. Tong J, Barany F, Cao W. Ligation reaction specificities of an NAD(+)-dependent DNA ligase from the hyperthermophile *Aquifex aeolicus*. Nucleic Acids Res 2000; 28:1447–1454.
47. Kvaratskhelia M, Wardleworth BN, Norman DG, White MF. A conserved nuclease domain in the archaeal Holliday junction resolving enzyme Hjc. J Biol Chem 2000; 275:25,540–25,546.
48. Yeh AP, Hu Y, Jenney FE Jr, Adams MW, Rees DC. Structures of the superoxide reductase from *Pyrococcus furiosus* in the oxidized and reduced states. Biochemistry 2000; 39:2499–2508.
49. Fiala G, Stetter KO. *Pyrococcus furiosus* sp nov represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch. Microbiol 1985; 145:56–61.
50. Hickey AJ, Conway de Macario E, Macario AJ. Transcription in the archaea: basal factors, regulation, and stress gene expression. Crit Rev Biochem Mol Biol 2002; 37:199–258.
51. Laksanalamai P, Maeder DL, Robb FT. Regulation and mechanism of action of the small heat shock protein from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 2001; 183: 5198–5202.
52. Kim KK, Kim R, Kim SH. Crystal structure of a small heat-shock protein. Nature 1998; 394: 595–599.
53. Laksanalamai P, Maeder DL, Jiemjit A, Bue Z, Robb FT. Oligomeric structures are necessary for cellular thermoadaption in the archaeal small heat shock protein. Extremophiles 2003; 7:79–83.
54. Jaenicke R, Bohm G. The stability of proteins in extreme environments. Curr Opin Struct Biol 1998; 8:738–748.
55. Robb FT, Maeder DL. Novel evolutionary histories and adaptive features of proteins from hyperthermophiles. Curr Opin Biotechnol 1998; 9:288–291.
56. Robb FT, Clark DS. Adaptation of proteins from hyperthermophiles to high pressure and high temperature. J Mol Microbiol Biotechnol 1999; 1:101–105.
57. Kaneshiro SM, Clark DS. Pressure effects on the composition and thermal behavior of lipids from the deep-sea thermophile *Methanococcus jannaschii*. J Bacteriol 1995; 177:3668–3672.
58. Sun MM, Tolliday N, Vetriani C, Robb FT, Clark DS. Pressure-induced thermostabilization of glutamate dehydrogenase from the hyperthermophile *Pyrococcus furiosus*. Protein Sci 1999; 8:1056–1063.
59. Sun MM, Caillot R, Mak G, Robb FT, Clark DS. Mechanism of pressure-induced thermostabilization of proteins: studies of glutamate dehydrogenases from the hyperthermophile *Thermococcus litoralis*. Protein Sci 2001; 10:1750–1757.
60. Cambillau C, Claverie JM. Structural and genomic correlates of hyperthermostability. J Biol Chem 2000; 275:32,383–32,386.
61. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc Natl Acad Sci USA 1999; 96:3578–3583.
62. Anderson DE, Becktel WJ, Dahlquist FW. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry 1990; 29:2403–2408.
63. Kumar S, Nussinov R. Salt bridge stability in monomeric proteins. J Mol Biol 1999; 293:1241–1255.
64. Lounnas V, Wade RC. Exceptionally stable salt bridges in cytochrome P450cam have functional roles. Biochemistry 1997; 36:5402–5417.

65. Elcock AH. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. J Mol Biol 1998; 284:489–502.
66. Yip KSP, Stillman TJ, Britton KL, et al. The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. Structure 1995; 3:1147–1115.
67. Vetriani C, Maeder DL, Tolliday N, et al. Protein thermostability above 100°C: a key role for ionic interactions. Proc Natl Acad Sci USA 1998; 95:12,300–12,305.
68. Consalvi V, Chiaraluce R, Giangiacomo L, et al. Thermal unfolding and conformational stability of the recombinant domain II of glutamate dehydrogenase from the hyperthermophile *Thermotoga maritima.* Protein Eng 2000; 13:501–507.
69. Lesley SA, Kuhn P, Godzik A, et al. Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. Proc Natl Acad Sci USA 2002; 99:11,664–11,669.
70. Wang W, Kim R, Jancarik J, Yokota H, Kim SH. Crystal structure of phosphoserine phosphatase from *Methanococcus jannaschii,* a hyperthermophile, at 1.8 A resolution. Structure (Camb) 2001; 10:9, 65–71.
71. Lee BI, Chang C, Cho SJ, et al. Crystal structure of the MJ0490 gene product of the hyperthermophilic archaebacterium *Methanococcus jannaschii,* a novel member of the lactate/malate family of dehydrogenases. J Mol Biol 2001; 307:1351–1362.
72. Lee DY, Ahn BY, Kim KS. A thioredoxin from the hyperthermophilic archaeon *Methanococcus jannaschii* has a glutaredoxin-like fold but thioredoxin-like activities. Biochemistry 2000; 39:6652–6659.
73. Wang H, Boisvert D, Kim KK, Kim R, Kim SH. Crystal structure of a fibrillarin homologue from *Methanococcus jannaschii,* a hyperthermophile, at 1.6 A resolution. EMBO J 2000; 19: 317–323.
74. Roy R, Adams MW. Tungsten-dependent aldehyde oxidoreductase: a new family of enzymes containing the pterin cofactor. Met Ions Biol Syst 2002; 39:673–697.
75. DiRuggiero J, Robb FT, Jagus R, et al. Cloning, characterization and in vitro expression of an extremely thermostable glutamate dehydrogenase from a novel hyperthermophilic Archeon, ES4. J Biol Chem 1993; 268:17,767–17,744.
76. Frankenberg RJ, Hsu TS, Yakota H, Kim R, Clark DS. Chemical denaturation and elevated folding temperatures are required for wild-type activity and stability of recombinant *Methanococcus jannaschii* 20S proteasome. Protein Sci 2001; 10:1887–1896.
77. Grottesi A, Ceruso MA, Colosimo A, Di Nola A. Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. Proteins 2002; 46:287–294.
78. Dairi T, Motohira Y, Kuzuyama T, Takahashi S, Itoh N, Seto H. Cloning of the gene encoding 3-hydroxy-3-methylglutaryl coenzyme A reductase from terpenoid antibiotic-producing *Streptomyces* strains. Mol Gen Genet 2000; 262:957–964.
79. Boucher Y, Huber H, L'Haridon S, Stetter KO, Doolittle WF. Bacterial origin for the isoprenoid biosynthesis enzyme HMG-CoA reductase of the archaeal orders Thermoplasmatales and Archaeoglobales. Mol Biol Evol 2001; 18:1378–1388.
80. Bochar DA, Stauffacher CV, Rodwell VW. Sequence comparisons reveal two classes of 3-hydroxy-3-methylglutaryl coenzyme A reductase. Mol Genet Metab 1999; 66:122–127.
81. Lecompte O, Ripp R, Puzos-Barbe V, et al. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res 2001; 11:981–993.
82. Maeder DL, Weiss RB, Dunn DM, et al. Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. Genetics 1999;

152:1299–1305.

83. Ettema T, van der Oost J, Huynen M. Modularity in the gain and loss of genes: applications for function prediction. Trends Genet 2001; 17:485–487.
84. DiRuggiero J, Dunn D, Maeder DL, et al. Evidence of recent lateral gene transfer among hyperthermophilic archaea. Mol Microbiol 2000; 38:684–693.
85. Kengen SW, de Bok FA, van Loo ND, Dijkema C, Stams AJ, de Vos WM. Evidence for the operation of a novel Embden–Meyerhof pathway that involves ADP-dependent kinases during sugar fermentation by *Pyrococcus furiosus*. J Biol Chem 1994; 269:17,537–17,541.
86. Tuininga JE, Verhees CH, van der Oost J, Kengen SW, Stams AJ, de Vos WM. Molecular and biochemical characterization of the ADP-dependent phosphofructokinase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem 1999; 274:21,023–21,028.
87. Kengen SW, Tuininga JE, de Bok FA, Stams AJ, de Vos WM. Purification and characterization of a novel ADP-dependent glucokinase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem 1995; 270:30,453–30,457.
88. Verhees CH, Tuininga JE, Kengen SW, Stams AJ, van der Oost J, de Vos WM. ADP-dependent phosphofructokinases in mesophilic and thermophilic methanogenic archaea. J Bacteriol 2001; 183:7145–7153.
89. Sakuraba H, Yoshioka I, Koga S, et al. ADP-dependent glucokinase/phosphofructokinase, a novel bifunctional enzyme from the hyperthermophilic archaeon *Methanococcus jannaschii*. J Biol Chem 2002; 277:12,495–12,498.
90. Ronimus RS, Kawarabayasi Y, Kikuchi H, Morgan HW. Cloning, expression and characterisation of a Family B ATP-dependent phosphofructokinase activity from the hyperthermophilic crenarchaeon *Aeropyrum pernix*. FEMS Microbiol Lett 2001; 202:85–90.
91. Mai X, Adams MW. Purification and characterization of two reversible and ADP-dependent acetyl coenzyme A synthetases from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 1996; 178:5897–5903.
92. Musfeldt M, Schonheit P. Novel type of ADP-forming acetyl coenzyme A synthetase in hyperthermophilic archaea: heterologous expression and characterization of isoenzymes from the sulfate reducer *Archaeoglobus fulgidus* and the methanogen *Methanococcus jannaschii*. J Bacteriol 2002; 184:636–644.
93. Silva PJ, van den Ban EC, Wassink H, et al. Enzymes of hydrogen metabolism in *Pyrococcus furiosus*. Eur J Biochem 2000; 267:6541–6551.
94. Kelly RM, Adams MW. Metabolism in hyperthermophilic microorganisms. Antonie Van Leeuwenhoek 1994; 66:247–270.
95. Adams MW, Holden JF, Menon AL, et al. Key role for sulfur in peptide metabolism and in regulation of three hydrogenases in the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 2001; 183:716–724.
96. Albers SV, van de Vossenberg JL, Driessen AJ, Konings WN. Adaptations of the archaeal cell membrane to heat stress. Front Biosci 2000; 5:D813–D820.
97. Albers SV, Driessen AM. Signal peptides of secreted proteins of the archaeon *Sulfolobus solfataricus*: a genomic survey. Arch Microbiol 2002; 177:209–216.
98. Koning SM, Albers SV, Konings WN, Driessen AJ. Sugar transport in (hyper)thermophilic archaea. Res Microbiol 2002; 153:61–67.
99. Kelly DJ, Thomas GH. The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. FEMS Microbiol Rev 2001; 25:405–424.
100. Albers SV, Van de Vossenberg JL, Driessen AJ, Konings WN. Bioenergetics and solute uptake under extreme conditions. Extremophiles 2001; 5:285–294.
101. Gonzalez JM, Kato C, Horikoshi K. *Thermococcus peptonophilus* sp nov, a fast-growing, extre-

mely thermophilic archaebacterium isolated from deep-sea hydrothermal vents. Arch Microbiol 1995; 164:159–164.

102. Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW. Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. Nucleic Acids Res 2000; 28:706–709.
103. Borges KM, Brummet S, Davis MF, et al. Survey of the genome of the hyperthermophile, *Pyrococcus furiosus*. Genome Sci Technol 1996; 1:37–46.
104. Gonzalez JM, Robb FT. Genetic analysis of the *Carboxydothermus hydrogenoformans* carbon monoxide dehydrogenase genes cooF and cooS(1). FEMS Microbiol Lett 2000; 191:243–247.
105. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science 1999; 283:220–221.
106. Korbel JO, Snel B, Huynen MA, Bork P. SHOT: a Web server for the construction of genome phylogenies. Trends Genet 2002; 18:158–162.
107. Breitung J, Borner G, Scholz S, Linder D, Stetter KO, Thauer RK. Salt dependence, kinetic properties and catalytic mechanism of *N*-formylmethanofuran:tetrahydromethanopterin formyltransferase from the extreme thermophile *Methanopyrus kandleri*. Eur J Biochem 1992; 210: 971–981.
108. Ermler U, Merckel M, Thauer R, Shima S. Formylmethanofuran:tetrahydromethanopterin formyltransferase from *Methanopyrus kandleri*—new insights into salt-dependence and thermostability. Structure 1997; 5:635–646.
109. Shima S, Herault DA, Berkessel A, Thauer RK. Activation and thermostabilization effects of cyclic 2, 3-diphosphoglycerate on enzymes from the hyperthermophilic *Methanopyrus kandleri*. Arch Microbiol 1998; 170:469–472.
110. Martusewitsch E, Sensen CW, Schleper C. High spontaneous mutation rate in the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by transposable elements. J Bacteriol 2000; 182:2574–2581.
111. Brugger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA. Mobile elements in archaeal genomes. FEMS Microbiol Lett 2002; 206:131–141.
112. She Q, Brugger K, Chen L. Archaeal integrative genetic elements and their impact on genome evolution. Res Microbiol 2002; 153:325–332.
113. Watanabe Y, Yokobori S, Inaba T, et al. Introns in protein-coding genes in Archaea. FEBS Lett 2002; 510:27–30.
114. Gabriel JL, Chong PL. Molecular modeling of archaebacterial bipolar tetraether lipid membranes. Chem Phys Lipids 2000; 105:193–200.
115. Schut GJ, Zhou J, Adams MW. DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus:* evidence for a new type of sulfur-reducing enzyme complex. J Bacteriol 2001; 183:7027–7036.
116. Shockley KR, Ward DE, Chhabra SR, Conners SB, Montero CI, Kelly RM. Heat shock response by the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl Environ Microbiol 2003; 69: 2365–2371.
117. Vierke G, Engelmann A, Hebbeln C, Thomm M. A novel archaeal transcriptional regulator of heat shock response. J Biol Chem 2002; October 14; Epub.
118. Dahlke I, Thomm M. A *Pyrococcus* homolog of the leucine-responsive regulatory protein, LrpA, inhibits transcription by abrogating RNA polymerase recruitment. Nucleic Acids Res 2002; 30:701–710.
119. Srinivasan G, James CM, Krzycki JA. Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science 2002; 296:1459–1462.
120. Ibba M, Soll D. Genetic code: introducing pyrrolysine. Curr Biol 2002; 12:R464–R466.

16

病原肠细菌基因组

Julian Parkhill and Nicholas R. Thomson

引言

肠杆菌科是包含许多种类细菌的一个组合，因为它们一般生活在动物肠道中（但并不绝对），而命名为肠细菌。肠细菌包括许多人类及家畜的重要病原体，并为微生物学研究提供了最永久的实验工具之一，如大肠杆菌 K12。因此，几十年来，科学家们已对肠细菌的生物学进行了深入研究[1]，很自然也用基因组学这一新兴科学对其进行了详尽研究。

肠细菌家族的核心是大肠杆菌（*Escherichia*）和沙门氏菌（*Salmonella*）两个种，本章主要介绍这两种菌。迄今为止，大肠杆菌 5 个菌株及沙门氏菌 2 个菌株公布了全序列，今后会有更多报道。然而，肠细菌还包括多种其他人类病原体，如鼠疫耶尔森氏菌（*Yersinia pestis*）、小肠结肠耶尔森氏菌（*Yersinia enterocolotica*）和肺炎克雷氏伯菌（*Klebsiella pneumonia*），还有植物病原体，如欧文氏菌（*Erwiniaceae*）、昆虫病原体，如光杆菌（*Photorhabdus*），以及昆虫共生生物，如巴克那氏菌（*Buchnera*）和 *Wigglesworthia*。已经或正在对这些病原体基因组进行测序（表 1），尽管大部分没有在这里直接讨论，但从大肠杆菌和沙门氏菌中得到的许多结果和知识与这些生物都有直接联系。

表 1　已完成或正在进行测序的肠细菌基因组

	大小/Mb	G+C/%	预测的基因数目	参考文献
完成				
蚜虫巴克纳氏菌 SG	0.641	25.3	545	[97]
蚜虫巴克纳氏菌 BP	0.616	25.3	504	[98]
蚜虫巴克纳氏菌 APS	0.641	26.3	564	[99]
大肠杆菌 K12	4.639	50.8	4289	[2]
大肠杆菌 O157:H7 (EPEC) EDL933	5.529	50.4	5349	[3]
大肠杆菌 O157:H7 (EPEC) RIMD (Sakai)	5.498	50.5	5361	[4]
大肠杆菌 CFT073(UPEC)	5.231	50.5	5379	[5]
肠沙门氏菌 Typhi CT18 血清变种	4.809	52.1	4599	[21]
肠沙门氏菌 Typhimurium LT2 血清变种	4.857	52.2	4489	[27]
弗氏志贺氏菌 2a	4.607	50.9	4434	[6]
Wigglesworthia brevipalpis	0.698	22.5	611	[100]
鼠疫耶尔森氏菌 CO-92	4.654	47.6	4012	[67]
鼠疫耶尔森氏菌 KIM	4.601	47.6	4198	[94]

续表

	大小/Mb	G+C/%	预测的基因数目	参考文献
未完成	资料来源(http://)			
胡萝卜软腐欧文氏菌	www.sanger.ac.uk/Projects/Microbes/			
菊欧文氏菌	www.ahabs.wisc.edu/~pernalab/			
大肠杆菌 042	www.sanger.ac.uk/Projects/Microbes/			
大肠杆菌 DH10B	www.hgsc.bcm.tmc.edu/microbial/			
大肠杆菌 E238/69	www.sanger.ac.uk/Projects/Microbes/			
大肠杆菌 K1	www.genome.wisc.edu/sequenceing.htm			
肺炎克雷氏伯菌	Genome.wustl.edu/Projects/bacterial/			
发光光杆菌	www.pasteur.fr/recherche/unites/gmp/			
Photorhabdus asymbiotica	www.sanger.ac.uk/Projects/microbes/			
乍得沙门氏菌	www.sanger.ac.uk/Projects/microbes/			
肠沙门氏菌亚利桑那血清变种	Genome.wustl.edu/Projects/bacterial/			
肠沙门氏菌 Choleraesuis 血清变种	www.salmonella.org/			
肠沙门氏菌 Diarizonae 血清变种	Genome.wustl.edu/Projects/bacterial/			
肠沙门氏菌 Dubin 血清变种	www.salmonella.org/			
肠沙门氏菌 Enteritidis LK5 血清变种	www.salmonella.org/			
肠沙门氏菌 Enteritidis PT4 血清变种	www.sanger.ac.uk/Projects/Microbes/			
肠沙门氏菌 Gallinarum 血清变种	www.sanger.ac.uk/Projects/Microbes/			
肠沙门氏菌 Paratyphi A 血清变种	Genome.wustl.edu/Projects/bacterial/			
肠沙门氏菌 Pullorum 血清变种	www.salmonella.org/			
肠沙门氏菌 Typhi Ty2 血清变种	www.genome.wisc.edu/sequenceing.htm			
肠沙门氏菌 Typhimurium DT104 血清变种	www.sanger.ac.uk/Projects/Microbes/			
肠沙门氏菌 Typhimurium SL1344 血清变种	www.sanger.ac.uk/Projects/Microbes/			
黏质沙雷氏菌	www.sanger.ac.uk/Projects/Microbes/			
痢疾志贺氏菌	www.sanger.ac.uk/Projects/Microbes/			
弗氏志贺氏菌 2*a*	www.genome.wisc.edu/sequenceing.htm			
索氏志贺氏菌	www.sanger.ac.uk/Projects/Microbes/			
小肠结肠炎耶尔森氏菌	www.sanger.ac.uk/Projects/Microbes/			
假结核耶尔森氏菌	www.bbrp.llnl.gov/bbrp/html/microbe.html			

注：数据来自 GOLD(2a;http://wit.integratedgenomics.com/GOLD/)和基因组研究所(http://www.tigr.org/tdb/mdb/mdbinprogress.html)。

将这些生物通过比较可以看出，它们有一个基本共线性基因组主干以及与之一致的核心系列基因及功能，同时，基因组中也出现了许多不同的点和区域。基因组之间的差异有几种独立却重叠的类型，这些将在下面的部分中讨论。这些差异包括大规模插入和删除（包括致病岛)、噬菌体整合、小规模插入和删除、点突变及染色体重排。

大规模插入和删除及致病岛概念

非致病大肠杆菌 K12 标准菌株 MG1655 的基因组序列于 1997 年[2]发表，此后，它成为与随后得到的肠细菌基因组序列进行比较的背景，还有两种基因型大肠杆菌已完成测序并公布了信息。此外，表 1 中列出了更多肠细菌的研究正在进行，如肠出血性大肠

杆菌（enterohemorrhagic *E. coli*，EHEC）O157：H7[3,4] 和尿道致病性大肠杆菌（uropathogenic *E. coli*，UPEC)[5]。关于弗氏志贺氏菌（*Shigella flexneri*）菌株 2a 基因组的信息已经公布[6]，现在可认为，志贺氏菌是大肠杆菌复合群的一类[7]，本章也这样划分。

1982 年首次分离到大肠杆菌 O157：H7，认为是引发大肠炎出血性痢疾和偶发性溶血性尿毒综合征（sporadically hemolytic uremic syndrome）的主要病因[8,9]。两个独立研究组已发表了菌株 EHEC 的基因组序列，Perna 等[3]发表了菌株 EDL933 的序列，它与最初记载的疾病暴发有关，该菌株已发表的序列包含两个物理缺口（4kb 及 54kb），为原噬菌体相关区域。Hayashi 等[4]发表了大肠杆菌 O157：H7 菌株 RIMD0509952 的完整序列，该菌株又称为 Sakai，是 1996 年日本大暴发此病时从病人体内分离的。

已测序的菌株 UPEC CFT073 是种类繁多肠外大肠杆菌的成员，UPEC 是引发新生儿脑膜炎/脓毒症（neonatal meningitis/sepsis）的病因及泌尿系统感染的主要原因。

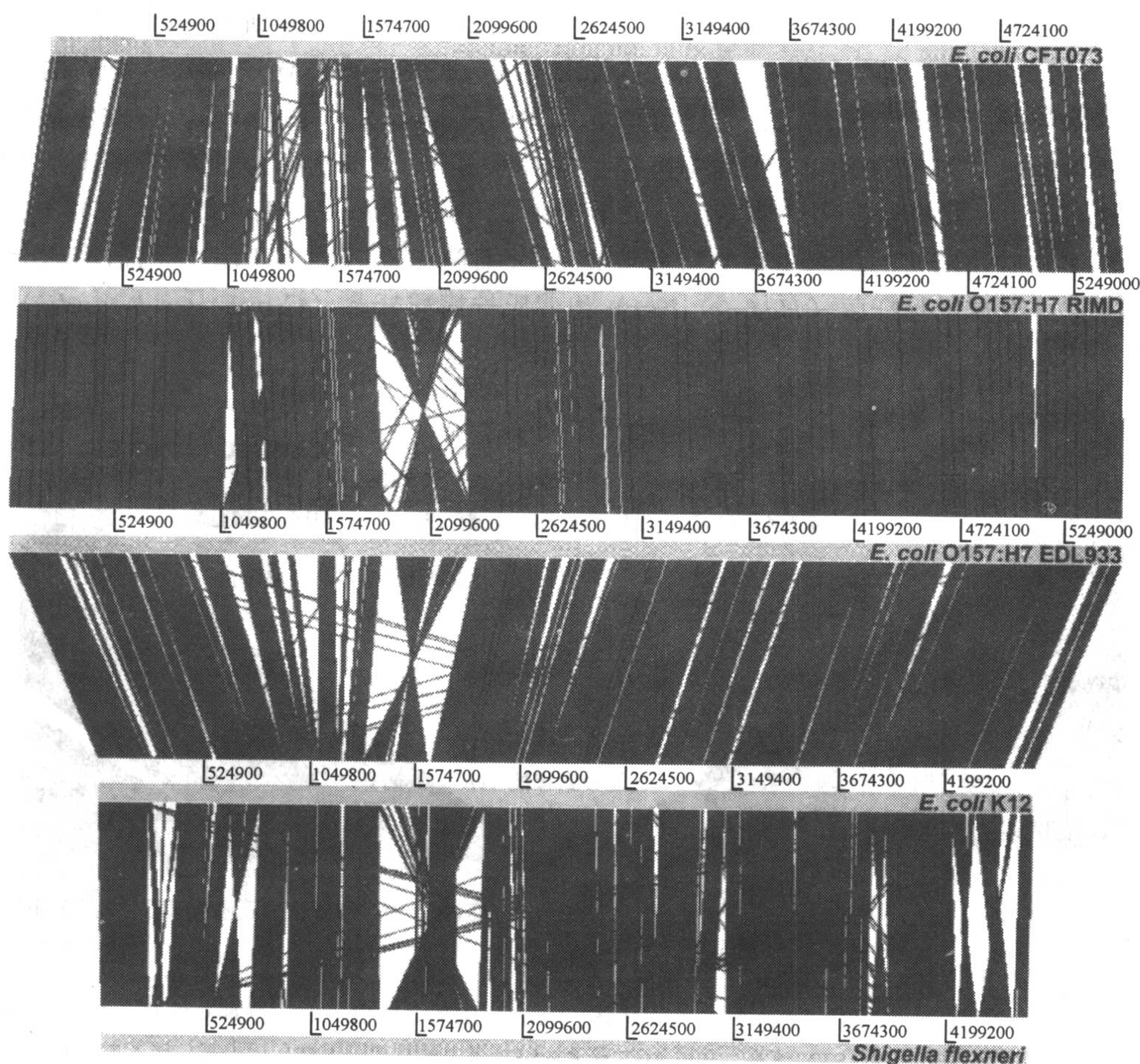


图 1　五个大肠杆菌菌株的全基因组比较。用 BLASTN 计算 DNA 匹配；展示方式利用 Artemis 比对工具（ACT；http://www.sanger.ac.uk/Software/ACT)。这些基因组从顶端开始依次为：*E. coli* CFT073（UPEC)、*E. coli* O157：H7 RIMD、*E. coli* O157：H7 EDL933、*E. coli* k12 和 *S. flexneri*。基因组间的灰色线条代表各个 BLASTN 匹配；为了显示比对的完整结构，较短或较微弱的 BLASTN 匹配被去除。

将所有致病大肠杆菌和非致病大肠杆菌 K12 原型进行比较发现，它们基因组的本质是共线性的，在序列及基因顺序上都显示出保守性（图 1）。与大肠杆菌 K12 相比，EHEC 和 UPEC 分别有 4.1Mb 和 3.9Mb 保守序列。据预测，这些保守序列中的编码基因显示，95%以上的序列相似性并定义为核心基因（core gene）。弗氏志贺氏菌基因组也得到以上类似的结果，在它的 4.6Mb 基因组中，有 3.9Mb 的序列与大肠杆菌 K12 相同[6]。

所有大肠杆菌基因组之间的比较表明：大肠杆菌菌株 K12、EHEC 和 UPEC 分别有约 4288、5400 和 5500 个预测的蛋白质编码序列，其中三个菌株共同的基因有 2996[5]个，编码这些核心基因的区域被看作为序列主干（backbone sequence）。

从这些比较中可以明显地看出，在整个序列主干内穿插着不同基因型特有的大型区域，此外，研究表明，这些特殊基因座有的在临床上引起的疾病中被分离出来，而在它们的无害近缘菌株中根本不存在[10]。

肠细胞损伤（enterocyte effacement）基因座（LEE，图 2）是研究得比较清楚的区域[11]，它在肠致病大肠杆菌（EPEC）中首次发现，EPEC 感染导致肠微绒毛的消除及细菌细胞与肠细胞的紧密黏附，它还会破坏细胞的完整结构并促使肌动蛋白发生聚合反应，肌动蛋白在附着的 EPEC 细胞中聚集形成杯状基座[12]，这称为附着拭除损伤（AE lesion）。导致附着拭除损伤的功能基因位于一个 35kb 区域——LEE，该区域中 G+C 含量异常，此后在发现能引起附着拭除损伤的所有细菌中都找到了 LEE，包括临床分离的 EHEC：哈夫尼肠杆菌（*Hafnia alvei*）和弗氏柠檬酸杆菌（*Citrobacter freundii*）菌株[11~16]。

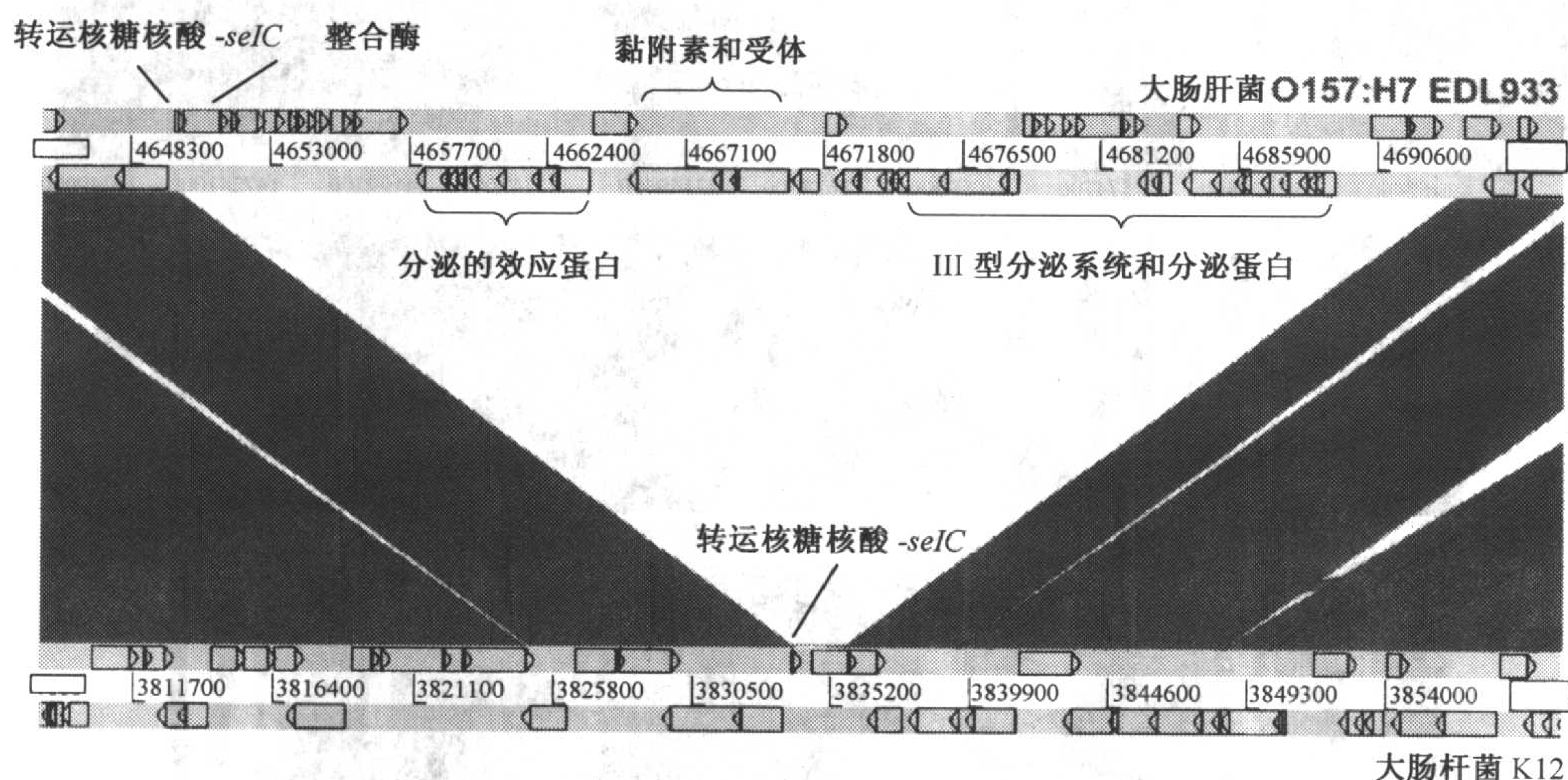


图 2　在 *E.coli* O157：H7 EDL933 中的 LEE 致病岛。*E.coli* O157：H7 EDL933（顶部）和 *E.coli* K12 间的 DNA：DNA 间的匹配用 BLASTN 计算，并用 ACT 展示。LEE 岛上的重要基因被标记出来。插入序列 DNA 的全长约为 45Kb，其中包括细菌噬菌体 933L（整合酶基因的右侧）。

更详细分析 LEE 显示，它位于 EPEC *selC* tRNA 基因的旁边，插入在与大肠杆菌 K12 同源的基因座中（图 2）。对该区域的序列分析表明：LEE 编码一种黏附素（in-

timin)、第三型分泌系统以及对致病至关重要的分泌型蛋白效应因子[11,17]，突变 LEE 会导致毒力减弱并丧失附着拭除损伤的能力[11,13]，此外，将 LEE 从 EPEC 引入大肠杆菌 K12，可使这一无害菌株也产生附着拭除损伤能力[18]。

许多与毒力有关的类似区域，在革兰氏阴性菌和革兰氏阳性菌中都有特征性描述（综述见参考文献［19］）。这引导出致病岛（pathogenicit island，PAI）概念以及描述其特征的法则。典型 PAI 一般插入到稳定 RNA 基因旁边，并有异常 G + C 含量，除了与毒力有关的功能外，这些区域经常携带编码转座酶及类似整合酶的基因，因而这些区域不稳定，能自己迁移[19,20]。在这里还需指出，在将致病岛与非致病岛比较时发现，PAI 分布受种系发生的限制，此外，与主干区域相比，PAI 携带很高比例的基因碎片或被破坏的基因[21]。

在很多一般性描述中，致病岛一词广泛用来描述许多基因座，有些基因座尽管拥有 PAI 的某些特征，却似乎与致病性无关。考虑到这点，PAI 的概念应扩展到包括那些代表种系特性和至今仍未发现与毒力有关的零散基因座[3,4]。

大肠杆菌所独有的区域/岛

除了主干序列，EHEC 和 UPEC 都附带一个约 1.3～1.4Mb 的独特序列。比较大肠杆菌菌株 K12 和 EHEC O157：H7 菌株 EDL933，独特区域划分为 K 岛和 O 岛[3]，分别对应于大肠杆菌 K12 和大肠杆菌 O157：H7 的独特区域，这些 K 岛和 O 岛与早先描述的 K 环（K12 特有）和 S 环（O157：H7 菌株特有）一致[4]。

采用 Hayashi 等的表示法[4]，大肠杆菌 O157：H7 有 296 个 S 环（>19kb），而大肠杆菌 K12 已发现了 325 个 K 环，S 环中约一半（48.2%）基因与噬菌体有关。在其余的基因中，33%功能未定，15%与毒力有关。在较大型 S 环中，有 4 个与毒力有关，编码 O157：H7 特有的纤毛蛋白，另外发现 5 个纤毛基因簇，在大肠杆菌 K12 中部分保守，能便于与菌株特异性位点结合，此外还发现了 14 个黏附素/侵染素，包括前面已描述 EHEC 特征的 LEE 致病岛中的黏附素和 Iha 黏附素[22]。

除了已位于 LEE 内的第三型分泌系统外，还发现了一种新第三型分泌系统。有趣的是，这种第三型分泌系统与肠沙门氏菌亚种鼠伤寒沙门氏菌（*S.typhimurium*）中的另一个致病岛，沙门氏菌致病岛 1（SPI-1）中的分泌系统更加相似（见下节）。除了前面描述过特征的噬菌体编码类志贺毒素（Shiga-like）（Stx）[23,24]和肠溶血素（enterohemolysin）外[25]，还发现了另两种毒素。据预测，其中一种毒素基因编码一个大型 RTX（repeats in structural toxin，结构毒素中的重复片段）家族蛋白（5292 aa），该蛋白位于 S 环，与那些促进分泌和活化的基因相邻。

Welch 等[5]采用了一种类似表示法描述 UPEC 特异岛（UI），UI 内编码基因包括 12 个不同纤毛系统，例如，已知有尿病原（uropathogen）特异性的两个 *pap* 操纵子，除此之外，在 UPEC 中已确认的还有其他几个纤毛系统，如 *yad* 分子伴侣介导系统，但也存在于 K12 和 EHEC 中。然而，就连这些很普遍的纤毛系统也显示出高度序列差异，表明它们与不同目标位点的相互作用。

还发现 UI 携带 7 个自身转运蛋白，一种 RTX 新毒素及 5 个 *fimE* 和 *fimB* 重组酶系统，这些系统都涉及与寄主的相互作用或转换（快速随机表型变化，与重组酶有关）。

总之，EHEC 的 S 环占全基因组 25.3%（1.393Mb），UI 则占 UPEC 基因组的 24.9%（1.303Mb），将 S 环和 UI 中的编码序列与大肠杆菌 K12 基因进行比较，Welch 等[5]发现，除了三种基因型所共有的 2996 个核心基因外，在 UPEC 中存在的 1827 个基因在大肠杆菌 K12 中却没有。在 EHEC 中具有的 1387 个基因在 K12 中却没有。大肠杆菌 K12 中有 585 个基因也是其他两种大肠杆菌所没有。有趣的是，在 UPEC 所特有的这 1827 个基因中，只有很小一部分（11%）存在于 EHEC 中。

对 S 环和 UI 核苷酸组成及密码子使用分析显示，它们有对主干序列（50.05%）非正常的 G+C 含量（47%～48%；这个数据排除了 EHEC 与噬菌体有关的基因），此外，稀有密码子在这些区域编码序列中占优势[3~5]。

这些观察结果与这样的假说一致，致病大肠杆菌基因型通过从比它小很多的非致病近缘菌获得外源 DNA 而进化来的，这些从侧向获得的 DNA，使不同基因型菌株定居在不同寄主的小生境，从而导致一系列不同的疾病。

大肠杆菌与沙门氏菌的比较

目前沙门氏菌有两个种的血清型超过了 2300 种，它们是肠沙门氏菌（*S. enterica*）和乍得沙门氏菌（*S. bongori*）[26]，所有沙门氏菌都有高度亲缘关系，它们之间的 DNA 相似性为 85%～95%。尽管有这样高的同质性，但不同血清型的病理和寄主范围有显著差异，例如，肠沙门氏菌肠亚种血清型伤寒沙门氏菌只能感染人并引起严重的伤寒热；而鼠沙门氏菌（*S. typhimurium*）引起人类肠炎、小鼠系统性感染，有很广的寄主范围[27]；肠沙门氏菌血清型鸡白痢沙门氏菌（*Pullorum*）主要感染鸡类，引起禽类伤寒[29]。

与大肠杆菌一样，沙门氏菌也拥有认为是侧向获得的沙门氏菌致病岛（salmonella pathogenicity island，SPI），是由 SPI-1[30,31]和 SPI-2[32,33]编码的基因产物，证实对感染过程的不同阶段都非常重要。这些岛都携带第三型分泌系统及相关分泌蛋白效应因子，SPI-1 赋予所有沙门氏菌侵染上皮细胞的能力，SPI-2 对感染过程的多个阶段都很重要，使沙门氏菌从肠道组织扩散到血液，并最终感染肝脏和脾脏中的巨噬细胞并在其中存活（综述见参考文献 [34]）。

与 LEE 和 PAI-1 一样，SPI-3（17kb）也插入到 *selC* tRNA 基因旁边[11,20]。SPI-3 携带基因 *mgtC*，*mgtC* 是细菌在巨噬细胞内生存和在吞噬泡内的低镁环境中生长所必需的[35]。

其他沙门氏菌的 SPI 编码第三型分泌效应蛋白、chaperone-usher 纤毛操纵子、Vi 抗原生物合成基因、一种 IVB 型菌毛操纵子及肠道致病性许多其他决定因子[21,36~39]。除了 SPI-7（Vi 致病岛）、SPI-8 和 SPI-10 外，伤寒沙门氏菌的大部分 SPI 同样也存在鼠伤寒沙门氏菌中。

SPI 有一些大肠杆菌 PAI 所没有的特征，SPI-1 是沙门氏菌所有成员所共有的，包括乍得沙门氏菌。肠沙门氏菌和乍得沙门氏菌是在沙门氏菌进化的早期阶段从一个共同祖先分化出来的，因此，可以认为，沙门氏菌从大肠杆菌中分化出来后，很快就获得了 SPI-1。肠沙门氏菌的每个亚种中都有 SPI-2，却在乍得沙门氏菌[40]中没有，所以它是在 SPI-1 之后才获得的，因此，与许多大肠杆菌 PAI 不同，它具有谱系特异性，而不是

菌株特异性。然而值得注意的是，在其他肠细菌中都检测不到 SPI-1 和 SPI-2[41]，真正 PAI 的分布也有同样的限制。

尽管文献里经常提到 PAI 迁移的本质，很少有实验证据直接支持这个观察结果。对此的解释是，整合后的 PAI 移动基因逐渐退化，使 PAI 位置固定。当然，有证据支持这个观点，因为许多推测的 PAI 携带整合酶或转座酶的假基因或残余序列。这方面最好的例子是，首先耶尔森氏菌有特征性描述的高致病岛（HPI）[42]，耶尔森氏菌 HPI 可以在岛内噬菌体整合酶基因（*int*）的基础上分裂成两个谱系：①小肠结肠耶尔森氏菌（*Y.enterocolitica*）1B 型；②鼠疫耶尔森氏菌（*Y.pestis*）和假结核耶尔森氏菌（*Y.pseudotuberculosis*）[43]。小肠结肠耶尔森氏菌 HPI *int* 基因有一个点突变，而鼠疫耶尔森氏菌和假结核耶尔森氏菌 HPI 内的相应基因完好。

HPI 是 35～43kb 的一个岛，携带耶尔森菌素（yersiniabactin）铁载体产生和吸收基因，以及推测与移动有关的基因，如 *int*，它的侧翼是对移动同样重要的 17bp 的正向重复。HPI 类似元件在肠细菌，包括大肠杆菌、克来伯氏菌、肠杆菌、柠檬酸杆菌[4,44,45]等中都有广泛分布，与许多原噬菌体一样，HPI 与 *asn*-tRNA 基因相邻，tRNA 基因是噬菌体整合入基因组的常见位点[46]，其原因可能是 tRNA 基因在属间高度保守并有许多拷贝，这些位点的整合，一般涉及噬菌体的 *attP* 位点和细菌基因组的整合位点 *attB* 之间同源序列位点的特异性重组。

耶尔森氏菌 HPI 和其他的 PAI 移动方式，可能与噬菌体移动方式类似。为了验证这点，Rakin 等[47]构建了以小型 HPI 为基础的质粒，它携带 *int* 基因、选择性标记和修饰过的 *attB* 位点（在 HPI 两侧 17bp 正向重复序列的基础上构建），这种构建能在大肠杆菌菌株 *recA* 中自由整合，此外，还能整合进多个 *asn*-tRNA 基因座[47,48]。

tRNA 基因位点是许多其他 PAI 和噬菌体的整合位点，在肠道细菌中利用最多的整合位点是 *selC* tRNA 基因座，该位点在产志贺氏毒素的大肠杆菌和 EPEC 中是 LEE PAI，在 UPEC 中是 PAI-1[5]，在产肠毒素的大肠杆菌（ETEC）中是一个新编码 Tia（一种黏附素）[49]，在弗氏志贺氏菌中是 SHI-2 基因（编码细菌铁嵌合蛋白的生物合成）[50]，在鼠伤寒沙门氏菌和伤寒沙门氏菌中则是 SPI-3[21,35]。合成 Stx 大肠杆菌的一些菌株本来就缺乏 LEE 岛，它们 *selC* tRNA 基因座上的 PAI 与蛋白酶解活动有关，因为岛内有丝氨酸蛋白水解酶[51]。相同的 PAI 不仅可以与并系同源（paralogous）tRNA 基因整合（如 HPI 所示；见以上讨论），也能整合在非同源 tRNA 基因座中，LEE 在大肠杆菌 *selC* tRNA 基因座中和 *pheU* tRNA 基因旁边都能找到[52,53]。

Welch 等[5]发表了 UPEC 与 EHEC 之间的全面比较结果，表明在基因组序列确认的 UI 中有 13 个整合在 tRNA 基因旁边，与之相比，EHEC 有 10 个特异岛位于 tRNA 基因旁边，此外，其中 9 个岛在两种基因型中都位于相同的 tRNA 基因旁，还有 10 个在 UPEC 和 EHEC 两个基因组中，有的基因座被不同物种的特异性岛占据，这些区域中至少有一个会整合到稳定的 RNA 基因 *ssrA* 中[4,5]。

然而，事实仍很复杂，许多岛并不是单次插入的结果，而是一系列连续插入的产物[54]。鼠伤寒沙门氏菌 SPI-2 是一个典型例子，SPI-2 可能由至少两次独立插入而产生。从整合位点 *valV*-tRNA 基因开始，SPI-2 起始的 25.3kb 序列，携带着编码第三型分泌装置和蛋白效应因子的基因，SPI-2 的这部分只存在沙门氏菌中而在乍得沙门氏菌

中不存在。紧接 SPI-2 的长 14.5kb 的一部分序列，在沙门氏菌和乍得沙门氏菌中都存在，它携带利用四硫磺酸盐基因，这个区域被认为是更加古老的一次插入序列的一部分[54]。对不同 UPEC 菌株的相同 UI 序列进行比较发现，这些 UI 显示出高度多样性[5]。

显然，PAI 在肠杆菌科（Enterobacteriaceae）许多致病成员的进化中起重要作用，这在肠细菌天然种群中也非常明显，它们的基因组显示出巨大的差异性。例如，大肠杆菌基因组在大小上有 1Mb 差异，波动在 4.5Mb 和 5.5Mb 之间[55,56]。在研究 PAI 进化选择时，Boyd 和 Hartl 观察到 PAI 和基因组大小相关，即携带 PAI 的菌株容易拥有更大的基因组[57]，说明 PAI 在很大程度上导致基因组在进化上的差异。大肠杆菌和沙门氏菌有些菌株的错配检测修复机制有缺陷，其副作用是提高了菌株重新组合和获得 DNA 的能力，使它们能接受大量外源 DNA[58]。近来，一些基因组研究项目发现，噬菌体对基因组大小的明显可变性也有显著影响。

噬菌体

在细菌基因组序列中，经常可以发现整合噬菌体或原噬菌体，噬菌体对细菌产生的影响不应该被低估。在大肠杆菌独特区域一节中曾提到，在 EHEC O157：H7 EDL933 菌株的 S 环中有近 50%与噬菌体有关（见第 5 章）。通过迄今已发表全部信息的分析，所有肠细菌基因组序列，包括原噬菌体的数量和种类已完全被揭示，并对了解细菌进化有戏剧性的影响。除了在 EHEC Sakai 菌株中检测到 18 种原噬菌体序列外，还检测到大肠杆菌 K12、UPEC 和弗氏志贺氏菌都携带多种原噬菌体或类似元件[2,3,5,6]。这些差异并不局限在不同种或属之间，对 EHEC O157：H7 EDL933 菌株和 Sakai 菌株之间基因组序列的比较显示，原噬菌体的种类与整合位点差异显著，即使是亲缘关系非常接近的噬菌体内部序列也表现出显著差异[4,59]。

原噬菌体的数量和种类是造成更多遗传差异的丰富源泉，除了自身复制所必需的基因，噬菌体经常携带一些其他基因，如防止细菌被其他噬菌体超感染的基因，*old* 和 *tin*[60,61]，一些由噬菌体携带进入细菌的非噬菌体基因能编码通过溶源转换（lysogenic conversion）增强细菌毒力（综述见参考文献[61a]）。

除了有 LEE PAI 和能诱发附着拭除损伤，Stx 产生也是 EHEC 的特征之一，志贺氏毒素是强大细胞毒素的一个家族，当进入真核细胞时，它能起糖基化酶的作用，并将 28S rRNA 切开，使核糖体失活，从而阻止蛋白质合成[62]。EHEC 能产生两种不同的 Stx，由不同的噬菌体编码：CP933V（Sp15）和 BP933W（Sp5）[3,4,63]。编码 *stx* 基因位于类似 lambda 的噬菌体中，它们的表达与原噬菌体诱导有关，在噬菌体晚期基因发生抑制后，在用化学试剂（如丝裂霉素 C 和抗生素）治疗腹泻时，要特别注意这一点，因为这有可能诱发此类毒素的合成[64]。

EHEC 基因组中由原噬菌体携带的其他决定因子，包括一个肠溶血素（enterohemolysin）*hly2*[65]、一组 tRNA 基因[4]以及与血清抗性有关的基因，如 *lom* 基因（详细论述见参考文献[66]）。

其他肠细菌，如伤寒沙门氏菌、鼠伤寒沙门氏菌和小肠结肠耶尔森氏菌也包含大量

原噬菌体[21,27,67]。沙门氏菌的主要毒力决定因子为SPI-1和SPI-2携带的第三型分泌系统和与它们相关的蛋白效应因子[68,69]。很多第三型分泌蛋白效应因子由原噬菌体基因组携带，其中包括分别由鼠伤寒沙门氏菌中类γ原噬菌体Gifsy1、Gifsy2携带的*sseI*和*gogB*（富含亮氨酸的YopM家族效应蛋白）[70,71]。与EHEC原噬菌体一样，这些沙门氏菌原噬菌体对细菌寄主的致病能力有巨大影响，例如，鼠伤寒沙门氏菌去除Gifsy2的菌株对模式鼠的毒力减弱100多倍[71,72]。

伤寒沙门氏菌和鼠伤寒沙门氏菌都携带类P2原噬菌体SopE[73]，像它的名字一样，SopE噬菌体携带*sopE*基因，其产物由SPI-1第三型分泌系统分泌并通过激活RhoGT-Pases来提高细菌进入细胞的效率[74]。与其他Gifsy1和Gifsy2噬菌体携带的效应因子蛋白一样，*sopE*位于SopE噬菌体的可变尾丝区域，对这一区域的详细分析显示，*sopE*区域为一个3kb的序列组件sopE moron[73]，可在亲缘和非亲缘噬菌体之间传递。据推测，这个机制可以便于噬菌体与细菌寄主之间的水平基因转移，由此避开了由噬菌体超感染免疫造成水平基因转移的障碍[73]。

小规模插入和删除

尽管前面讨论大型PAI是解释菌株不同表型的基础是无可非议的，但当研究这组生物的总体基因组时，必须对其他类型的差异加以考虑。

大肠杆菌K12和大肠杆菌O157：H7[3]，以及伤寒沙门氏菌和鼠伤寒沙门氏菌之间的比较[21]，说明它们的基因组中除了前面讨论的大型岛外，还存在许多小的差异。事实上，对伤寒沙门氏菌与鼠伤寒沙门氏菌的比较，以及对伤寒沙门氏菌与大肠杆菌K12的比较清楚地显示，每对生物间多数插入或删除的规模很小，对伤寒沙门氏菌与鼠伤寒沙门氏菌之间单独插入或删除数据的比较显示，有145次事件只涉及10个或以下的基因，12次事件涉及20个或以上的基因。伤寒沙门氏菌与大肠杆菌K12之间的比较显示，有504次事件涉及10个或以下的基因，而只有25次事件涉及20个或以上的基因（图3A）。就算考虑到大型岛的每次插入或删除涉及较多的基因，很显然，由涉及10个或以下基因和涉及20个或以上基因的插入或删除，所造成种特异性基因的数量几乎相等（图3B）。在伤寒沙门氏菌与鼠伤寒沙门氏菌的比较中，涉及10个或以下基因的插入或删除导致了377个基因差异，而涉及20个或以上基因的事件导致了631个基因差异。同样地，在伤寒沙门氏菌与大肠杆菌K12的比较中，前者为1287，后者为1019（图3B）。

从讨论中可以明确看出，小规模岛的获得或交换对生物总体表型很重要，有些例子强化了这个观点。SspH2由*SPI-2*编码，由沙门氏菌第三型分泌系统分泌的富含亮氨酸蛋白；在伤寒沙门氏菌中，SspH2是由一个与大肠杆菌K12同源的区域编码，该区域包含噬菌体的一个完整基因和三个假基因；鼠伤寒沙门氏菌中相应的岛包含其他来源的噬菌体基因，说明该区域都是前噬菌体的残余。在另外情况下，鼠伤寒沙门氏菌的*envF*基因（编码一个预测的脂蛋白），在伤寒沙门氏菌中被一个与空肠弯杆菌（*Campylobacter jejuni*）毒素亚基CdtB和百日咳博德氏菌（*Bordetella pertussis*）毒素亚基PtxA和PtxB有远缘关系的5个基因模块所取代，在其他生物的比较中也发现许多这样的例

子。

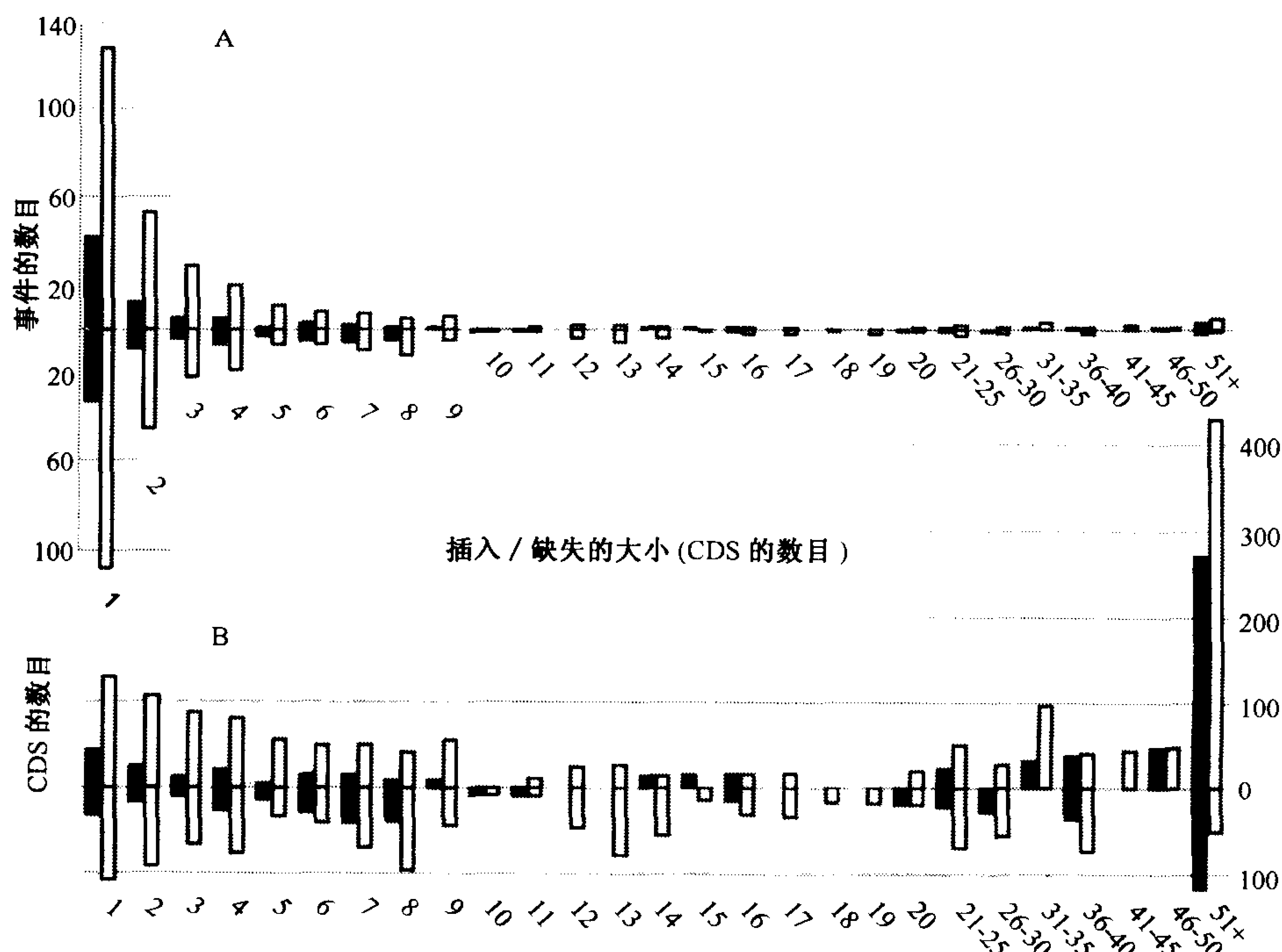


图 3　伤寒沙门氏菌与鼠伤寒沙门氏菌/大肠杆菌之间的插入和删除事件。X 轴代表按编码序列数目统计的插入/删除事件的多少。轴上方柱代表伤寒沙门氏菌与鼠伤寒沙门氏菌（灰色）和大肠杆菌（白色）的插入事件。轴下方柱代表伤寒沙门氏菌与鼠伤寒沙门氏菌（灰色）和大肠杆菌（白色）的删除事件。A 图中 Y 轴代表插入/删除事件的数目；B 图中 Y 轴代表插入/删除事件中编码序列的总数。

多数情况下，没有证据表明存在能使这些岛自己移动的基因，在理论上，这些是谱系特异基因的缺失，这几组基因在菌株或菌种共同祖先中存在。然而，如果事实果真如此，那么始祖染色体就必须比现存成员大得多，但这似乎是不可能的。更有可能的是，这种小规模岛确实在种内成员间交换，并构成该种基因库的一部分，显而易见，一旦这些岛被种内的某个成员获得，它们可以通过普通转导机制很方便地在种内成员间交换[1]，接着通过基本相同侧翼基因之间的同源重组整合进入染色体。

这种基因交换机制也会被非直系同源基因（nonorthologous gene）取代，其取代方式涉及染色体主干部分相同区域中相关基因的交换，一个特殊例子是脑膜炎奈瑟氏球菌（*Neisseria meninyitidis*）[75]和肺炎链球菌（*Streptococcus pneumoniae*）[76,77]的荚膜转换，它们对应荚膜多糖生物合成的系列基因，出现在染色体的相同区域并与保守基因相连。荚膜转换的机制可能涉及到通过侧翼基因介导染色体和外源 DNA 之间的同源重组，来进行特异多糖基因的取代。

这种基因转换现象更普遍，在大肠杆菌和沙门氏菌众多的 chaperone-usher 纤毛系统中都可能存在，这些系统在大肠杆菌[3]和肠沙门氏菌[78]中都存在但数量不定；Perna 等[3]指出，这些纤毛操纵子在大肠杆菌 K12 和大肠杆菌 O157：H7 中是最易变的序列，

并且两个基因组之间最易变的基因是直系同源基因 *yadC*，它编码这些操纵子中一个纤毛亚基；在伤寒沙门氏菌和大肠杆菌 K12 之间的比较也证实了这一观察结果。图 4 显示大肠杆菌中包含 *yadC* 操纵子和伤寒沙门氏菌中明显同源的 *sta* 纤毛操纵子，它们位于染色体同样的区域中，并且侧翼基因显示出高度相似性，这与纤毛基因自身较低保守性形成鲜明对比，这可能表明，在选择压力下（可能来自寄主免疫系统）序列迅速分化。此外，有类似但非直系同源基因在染色体同一位置通过侧翼保守基因和外源 DNA 间进行同源重组交换。有趣的是，尽管另一个 chaperone-usher 系统就位于下游 4kb 处，这个操纵子在鼠伤寒沙门氏菌中并不存在（图 4），这进一步阐明了功能性小岛的交换。

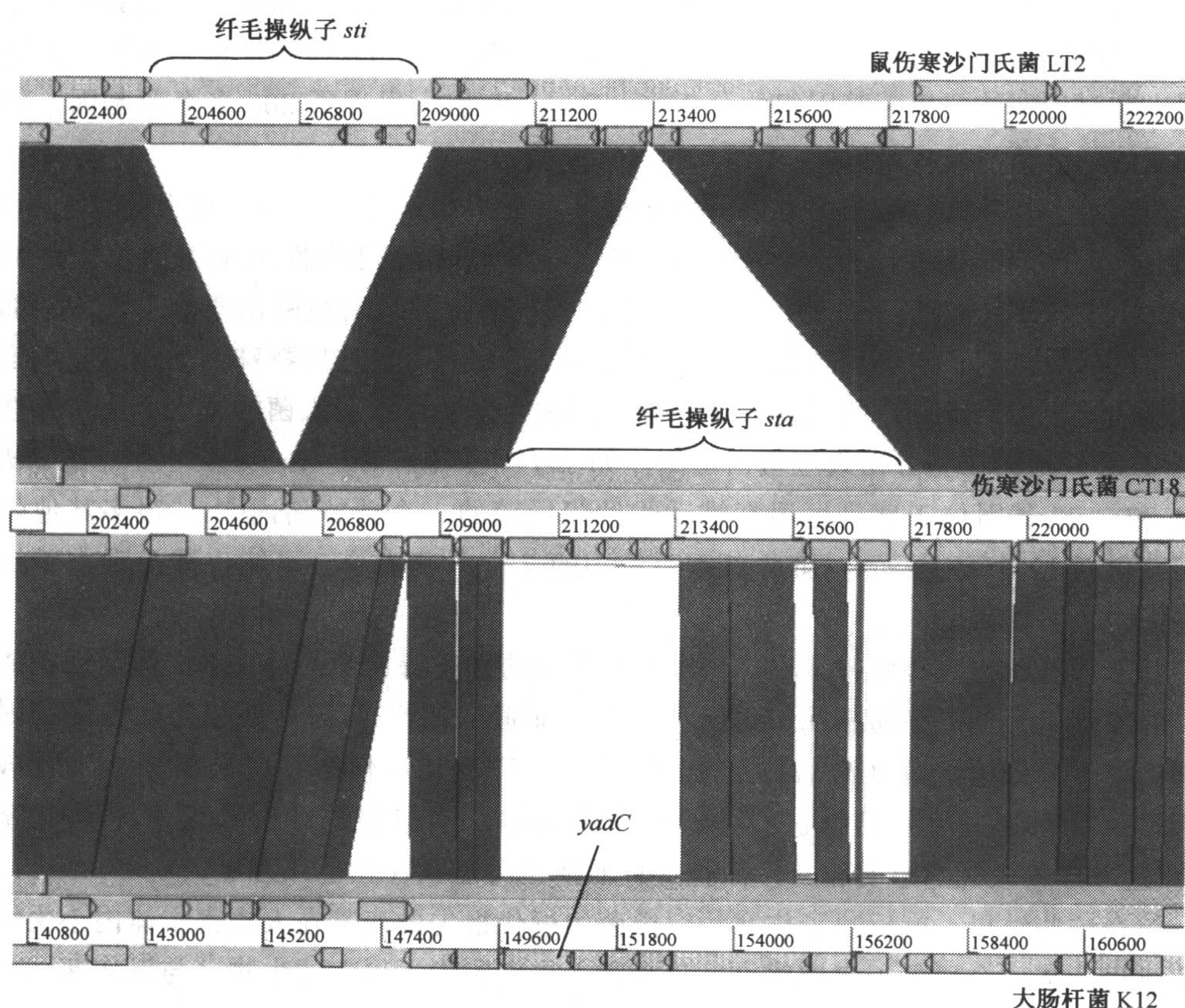


图 4 沙门氏菌和大肠杆菌的 chaperone-usher 纤毛操纵子。在鼠伤寒沙门氏菌 LT2、伤寒沙门氏菌 CT18 和大肠杆菌 K12 菌株间用 BLASTN 进行 DNA 序列相似性匹配，并用 ACT 进行展示。基因组之间灰色柱代表单个 BLASTN 匹配。伤寒沙门氏菌和鼠伤寒沙门氏菌中的纤毛操纵子 *sti* 和 *sta*，以及大肠杆菌中相应操纵子的一部分—*yadC* 基因均被标示。

点突变和假基因

从肠细菌基因组一些研究项目中，发现最令人感到惊奇的结果之一是，某些种/菌株包含很大量的假基因，即由于终止密码子、移码、内部缺失或插入序列（IS）的存

在，而导致无法翻译的基因。这已成为许多争论的起因，尤其是违背了公认的一种假说，即细菌基因组是不包含“垃圾 DNA”的一个高效率系统。

毫无疑问，某种特殊机制通过程序化核糖体移码，对终止密码子进行通读（read-through）从而纠正移码突变[1]。然而，在很多情况下，特殊表型与这些基因组突变之间存在某种联系，另一方面，比较基因组学的发展使人们对无功能性的预测有了更多的信心。假定两个近缘菌株一个有突变而另一个没有，那么前一个菌株能在如此短的进化时间内，既获得突变又获得特定抑制机制的可能性有多大？假定对失活基因的预测正确，那么生物学及进化论能从中推测出什么结论呢？

在大肠杆菌与沙门氏菌比较一节中曾提到，引发伤寒热的伤寒沙门氏菌具有寄主限制性，只能感染人类寄主，但引发比较温和人类疾病的鼠伤寒沙门氏菌则有较广的寄主范围，在对伤寒沙门氏菌基因组的分析中，Parkhill 等[21]发现了 200 多个假基因，但在鼠伤寒沙门氏菌只发现了约 39 个[27]。

很明显，伤寒沙门氏菌中的假基因在整个基因组中不随机分布，与大肠杆菌相比，它们在伤寒沙门氏菌特有基因中存在的比例很大（伤寒沙门氏菌 33％基因位于独特区域，却有 59％假基因位于独特区域）。伤寒沙门氏菌的假基因在鼠伤寒沙门氏菌中都有完整对应拷贝，它们与毒力以及与寄主间的相互作用有关。这些特殊例子包括：①富含亮氨酸的重复蛋白 SlrP，通过第三型系统分泌并与鼠伤寒沙门氏菌的寄主范围特异性有关[79]；②其他第三型分泌效应蛋白基因，如 *sseJ*[80]、*sopE2*[81]和 *sopA*[82,83]；③*shdA*、*ratA* 和 *sivH* 基因位于感染恒温脊椎动物沙门氏菌的一个独特岛内[84]。很多其他失活基因，包括 12 个 chaperone-usher 纤毛系统中的 7 个系统成分，可能也涉及毒力以及与寄主间的相互作用。

基于假基因的这种分布状况，伤寒沙门氏菌的寄主特异性，是由于必需基因功能失活，而导致与其他寄主间相互作用能力的丧失而造成的[21]。与其他包含几种假基因的生物，如麻风分枝杆菌（*Mycobacterium leprae*）[85]相比，伤寒沙门氏菌中大多数假基因都是由单突变引起，表明它们是近期才失活的，这与伤寒沙门氏菌看来是克隆化的[86]这一广为人知的事实相符，而且这种血清型可能只存在了几万年[86a]。

综合这些结果，就反映了伤寒沙门氏菌的近代祖先改变其在人类寄主体内小生境的进化过程，从一个局限于定域感染且只在肠上皮细胞附近侵入的生物（与鼠伤寒沙门氏菌类似），进化成能侵入人类寄主深层组织的生物，这种生境的变化可能涉及一个小种群，导致了进化瓶颈，通过遗传漂变增加变异[87]。

有人提出了另一个新进化肠细菌鼠疫耶尔森氏菌也有相似的进化过程，这种生物是最近从肠细菌假结核耶尔森氏菌的粪便传播途径，转变为能用跳蚤作为载体而进行系统性感染的生物[88,89]。这种生境的改变又一次伴随假基因的产生，并且涉及毒力和与寄主相互作用的基因在一系列失活基因中比例很高[67]。

更深入的例子是大肠杆菌的一个成员弗氏志贺氏菌 2a（预测有超过 250 个假基因），也局限于人类寄主[6]。

这些生物清楚地证明，肠细菌进化是与基因丢失和基因获取都有关的一个过程，在这个进化中缺失的基因残余很容易检测出来。

重排：基因组的完整性和可变性

在大肠杆菌和沙门氏菌最初完成染色体图谱的时候，观察到许多标记位于两个染色体的相似位置，而且两个基因组之间存在整体共线性[90,91]，现在的完整基因组序列充分支持了这个最初的观察结果。图1显示了已测序大肠杆菌4个菌株的基因组，除大肠杆菌O157：H7 EDL933在复制末端附近有440Kb的倒位外，基本上是完全共线性的[3]，因为它们都属同一个种，这种现象并不奇怪。

令人惊奇的是，这个共线性还延伸到沙门氏菌基因组（图5）[21,27]，仅仅是复制末端附近的一个倒位序列，就导致了鼠伤寒沙门氏菌与大肠杆菌K12全部基因顺序的不同。这些在复制起点和末端附近交叉的倒位序列，是亲缘细菌染色体重排的最普遍形式[92,93]，据推测是由复制叉间的直接重组所造成[92]，或是因为只有复制叉附近DNA被解包装，因而可以重组[94]。

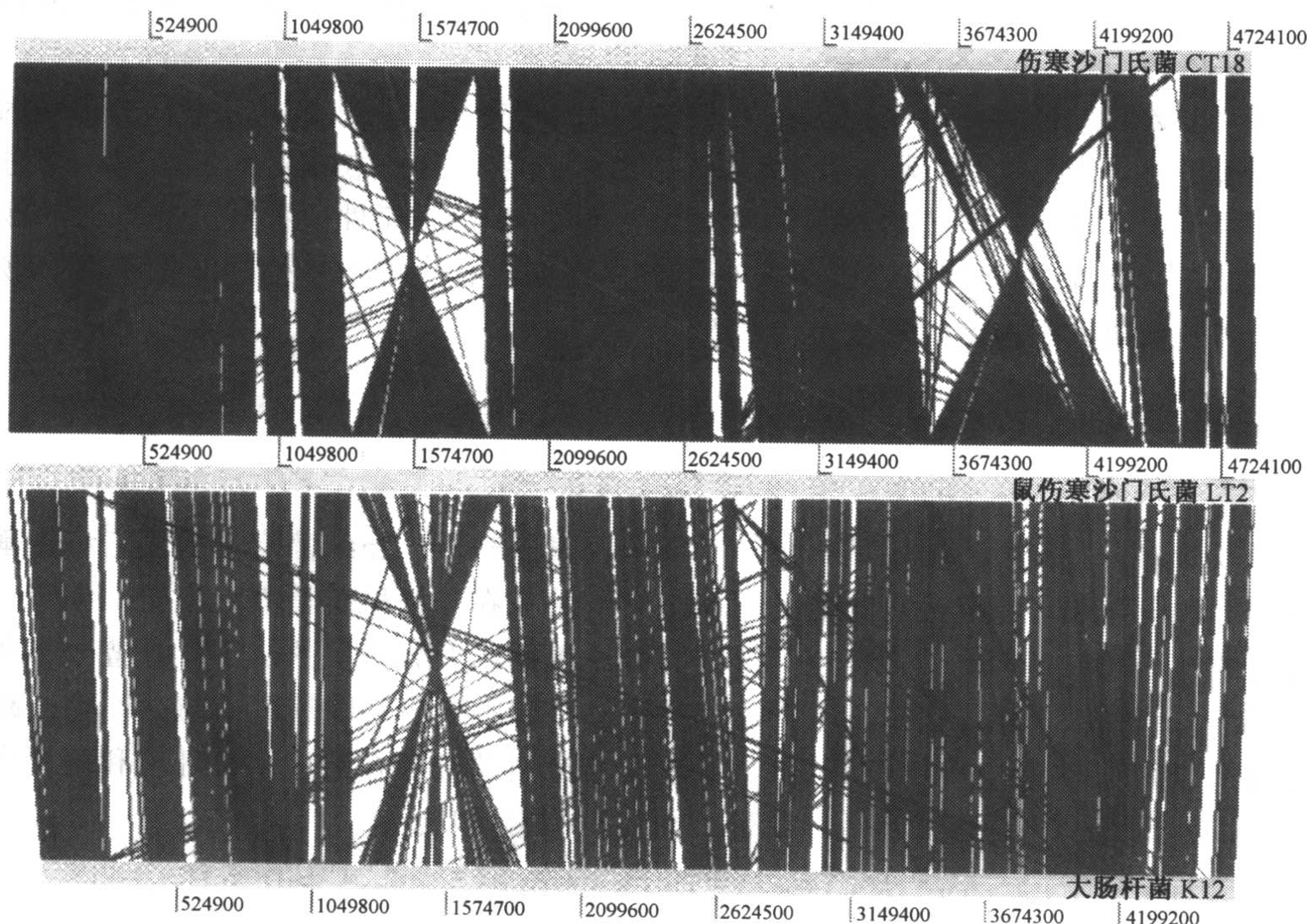


图5 沙门氏菌和大肠杆菌的全局比较。伤寒沙门氏菌CT18、鼠伤寒沙门氏菌LT2和大肠杆菌K12菌株间用BLASTN进行DNA序列相似性匹配，并用ACT进行展示。基因组之间灰色柱代表单个BLASTN匹配。较短和较弱的BLASTN匹配被除去以展示整体比较结果。

对整体基因序列保守性的两种可能解释是，这些基因组中几乎没有出现重组，因而很稳定，或者大肠杆菌和沙门氏菌特定基因顺序是一种选择的结果，并保持至今。缺少重组这种解释看来不合理，因为这些生物中重组非常频繁，有足够证据表明，在这些生物与更远的亲缘菌，如耶尔森氏菌之间也有基因重排[67,94]。此外，从基因组比较的细

节可以清楚看出，这些生物发生过染色体重组。

即使是不断重组，保守基因的顺序仍然通过选择而保持至今，如果确实如此，那么就有必要弄清这个选择的基础。有几种可能性：①肠细菌普遍都非常依赖功能相关基因的共调节，它们通过将功能相关基因聚集在共转录操纵子中实现这个目的，这有可能对操纵子内部重组产生限制。②在生长迅速的细菌中，往往在细胞分裂前很早染色体复制就开始了，这样，就使最靠近起点的基因比那些靠近末端的基因，在每个细胞中的拷贝数要多。这种基因剂量效应很可能对细胞特定表达水平产生影响，生物应该已能很好地适应并利用这个效应。③基因转录方向多与复制叉移动方向一致，以免复制和转录机制之间发生冲突，重组同样能破坏这个过程。

其他可能性，包括基因位置和取向上的选择和突变压力[95,96]，其中，某些因素或所有因素可能同时作用。有一点很有趣，在大肠杆菌和志贺氏菌的比较中，没有发现这个保守基因顺序。很显然，弗氏志贺氏菌的许多倒位和易位序列，不位于复制的起点或末端附近[6]（图 1)，所有这些都是由数量众多完全相同的 IS 元件间的重组造成，这些 IS 元件散布在弗氏志贺氏菌基因组中，包含 300 多个 IS 元件，这相当于大肠杆菌其他菌株的 7 倍多。

在大多数肠细菌中无论保持基因顺序和方向的选择压力是什么，它们都会被基因组内这个水平 IS 元件的扩展及随后的大规模重组所超越，这一点在对肠细菌另一个分支的观察中再次被证实。小肠结肠耶尔森氏菌经历过相似 IS 元件的扩展，在近缘小肠结肠耶尔森氏菌菌株[94]间的比较说明重组发生在近期，有些重组事件甚至发生在小肠结肠耶尔森氏菌单细胞菌落的形成过程中[67]。

结论

可以看出，正如所列举大肠杆菌和沙门氏菌家族一样，肠细菌有能力使它们的基因组达到一种既显著稳定又高度可变的平衡。它们共有在一条稳定基因组主干上编码一系列保守的核心功能基因。然而，有许多与特异性基因组变异有关的机制，重叠分布在这条主干上，包括大规模基因获取（PAI 和噬菌体）和小规模基因获取，与由切除、缺失和突变造成的基因缺失保持平衡，这些机制使这个科的细菌在适应多种环境和病原生境方面，以及迅速进化并得以存活和繁殖方面显得非常成功。

致谢

我们感谢 Stephen Bentley 和 Matt Holden 对手稿的严格审阅。

（江　炅　译）

参考文献

1. Neidhardt FC, Curtiss R. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. 2nd ed. Washington, DC: ASM Press, 1996.
2. Blattner FR, Plunkett G, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997 ; 277 :1453–1474.
2a. Bernal A, Ear U, Kyrpides N. Genome Online Databases (GOLD): a monitor of genome projects worldwide. Nuc Acids Res 2001; 29:126–127.
3. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia* coli O157:H7. Nature 2001; 409:529–533.
4. Hayashi T, Makino K, Ohnishi M, et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 2001; 8:11–22.
5. Welch RA, Burland V, Plunkett G 3rd, et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 2002; 99: 17,020–17,024.
6. Jin Q, Yuan Z, Xu J, et al. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res 2002; 30:4432–4441.
7. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci USA 2000; 97:10,567–10,572.
8. Riley LW, Remis RS, Helgerson SD, et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N Engl J Med 1983; 308:681–685.
9. Karmali MA, Petric M, Lim C, Fleming PC, Steele BT. *Escherichia coli* cytotoxin, haemolytic-uraemic syndrome, and haemorrhagic colitis. Lancet 1983; 2:1299–1300.
10. Knapp S, Hacker J, Jarchau T, Goebel W. Large, unstable inserts in the chromosome affect virulence properties of uropathogenic *Escherichia coli* O6 strain 536. J Bacteriol 1986; 168: 22–30.
11. McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. Proc Natl Acad Sci USA 1995; 92: 1664–1668.
12. Levine MM. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. J Infect Dis 1987; 155:377–389.
13. Jarvis KG, Giron JA, Jerse AE, McDaniel TK, Donnenberg MS, Kaper JB. Enteropathogenic *Escherichia coli* contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion formation. Proc Natl Acad Sci USA 1995; 92: 7996–8000.
14. Frankel G, Candy DC, Everest P, Dougan G. Characterization of the C-terminal domains of intimin-like proteins of enteropathogenic and enterohemorrhagic *Escherichia coli*, *Citrobacter freundii*, and *Hafnia alvei*. Infect Immun 1994; 62:1835–1842.
15. Donnenberg MS, Yu J, Kaper JB. A second chromosomal gene necessary for intimate attachment of enteropathogenic *Escherichia coli* to epithelial cells. J Bacteriol 1993; 175:4670–4680.
16. Schauer DB, Falkow S. Attaching and effacing locus of a *Citrobacter freundii* biotype that causes transmissible murine colonic hyperplasia. Infect Immun 1993; 61:2486–2492.
17. Kenny B, Finlay BB. Protein secretion by enteropathogenic *Escherichia coli* is essential for

transducing signals to epithelial cells. Proc Natl Acad Sci USA 1995; 92:7991–7995.

18. McDaniel TK, Kaper JB. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. Mol Microbiol 1997; 23: 399–407.
19. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 1997; 23:1089–1097.
20. Blum G, Ott M, Lischewski A, et al. Excision of large DNA regions termed pathogenicity islands from tRNA- specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. Infect Immun 1994; 62:606–614.
21. Parkhill J, Dougan G, James KD, et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature 2001; 413:848–852.
22. Tarr PI, Bilge SS, Vary JC Jr, et al. Iha: a novel *Escherichia coli* O157:H7 adherence-conferring molecule encoded on a recently acquired chromosomal island of conserved structure. Infect Immun 2000; 68:1400–1407.
23. O'Brien AD, Marques LR, Kerry CF, Newland JW, Holmes RK. Shiga-like toxin converting phage of enterohemorrhagic *Escherichia coli* strain 933. Microb Pathog 1989; 6:381–390.
24. O'Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB. Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. Science 1984; 226:694–696.
25. Schmidt H, Beutin L, Karch H. Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL 933. Infect Immun 1995; 63:1055–1061.
26. Boyd EF, Wang FS, Whittam TS, Selander RK. Molecular genetic relationships of the salmonellae. Appl Environ Microbiol 1996; 62:804–808.
27. McClelland M, Sanderson KE, Spieth J, et al. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 2001; 413:852–856.
28. Reeves P, Stevenson G. Cloning and nucleotide sequence of the *Salmonella typhimurium* LT2 gnd gene and its homology with the corresponding sequence of *Escherichia coli* K12. Mol Gen Genet 1989; 217:182–184.
29. Shivaprasad HL. Fowl typhoid and pullorum disease. Rev Sci Tech 2000; 19:405–424.
30. Mills DM, Bajaj V, Lee CA. A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome. Mol Microbiol 1995; 15:749–759.
31. Galan JE. Molecular genetic bases of *Salmonella* entry into host cells. Mol Microbiol 1996; 20:263–271.
32. Shea JE, Hensel M, Gleeson C, Holden DW. Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*. Proc Natl Acad Sci USA 1996; 93:2593–2597.
33. Ochman H, Soncini FC, Solomon F, Groisman E. A. Identification of a pathogenicity island required for *Salmonella* survival in host cells. Proc Natl Acad Sci USA 1996; 93:7800–7804.
34. Kingsley RA, Baumler AJ. Pathogenicity islands and host adaptation of *Salmonella* serovars. Curr Top Microbiol Immunol 2002; 264:67–87.
35. Blanc-Potard AB, Groisman EA. The *Salmonella selC* locus contains a pathogenicity island mediating intramacrophage survival. EMBO J 1997; 16:5376–5385.
36. Wood MW, Jones MA, Watson PR, Hedges S, Wallis TS, Galyov EE. Identification of a pathogenicity island required for *Salmonella* enteropathogenicity. Mol Microbiol 1998; 29:883–891.
37. Galyov EE, Wood MW, Rosqvist R, et al. A secreted effector protein of *Salmonella dublin* is translocated into eukaryotic cells and mediates inflammation and fluid secretion in infected

ileal mucosa. Mol Microbiol 1997; 25:903–912.

38. Hashimoto Y, Li N, Yokoyama H, Ezaki T. Complete nucleotide sequence and molecular characterization of ViaB region encoding Vi antigen in *Salmonella typhi*. J Bacteriol 1993; 175: 4456–4465.
39. Zhang XL, Tsui IS, Yip CM, et al. *Salmonella enterica* serovar typhi uses type IVB pili to enter human intestinal epithelial cells. Infect Immun 2000; 68:3067–3073.
40. Hensel M, Shea JE, Baumler AJ, Gleeson C, Blattner F, Holden DW. Analysis of the boundaries of *Salmonella* pathogenicity island 2 and the corresponding chromosomal region of *Escherichia coli* K-12. J Bacteriol 1997; 179:1105–1111.
41. Lee CA. Pathogenicity islands and the evolution of bacterial pathogens. Infect Agents Dis 1996; 5:1–7.
42. Buchrieser C, Prentice M, Carniel E. The 102-kilobase unstable region of *Yersinia pestis* comprises a high-pathogenicity island linked to a pigmentation segment which undergoes internal rearrangement. J Bacteriol 1998; 180:2321–2329.
43. Rakin A, Urbitsch P, Heesemann J. Evidence for two evolutionary lineages of highly pathogenic *Yersinia* species. J Bacteriol 1995; 177:2292–2298.
44. Schubert S, Cuenca S, Fischer D, Heesemann J. High-pathogenicity island of *Yersinia pestis* in enterobacteriaceae isolated from blood cultures and urine samples: prevalence and functional expression. J Infect Dis 2000; 182:1268–1271.
45. Bach S, de Almeida A, Carniel E. The *Yersinia* high-pathogenicity island is present in different members of the family Enterobacteriaceae. FEMS Microbiol Lett 2000; 183:289–294.
46. Reiter WD, Palm P, Yeats S. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res 1989; 17:1907–1914.
47. Rakin A, Noelting C, Schropp P, Heesemann J. Integrative module of the high-pathogenicity island of *Yersinia*. Mol Microbiol 2001; 39:407–415.
48. Hare JM, Wagner AK, McDonough KA. Independent acquisition and insertion into different chromosomal locations of the same pathogenicity island in *Yersinia pestis* and *Yersinia pseudotuberculosis*. Mol Microbiol 1999; 31:291–303.
49. Fleckenstein JM, Lindler LE, Elsinghorst EA, Dale JB. Identification of a gene within a pathogenicity island of enterotoxigenic *Escherichia coli* H10407 required for maximal secretion of the heat-labile enterotoxin. Infect Immun 2000; 68:2766–2774.
50. Moss JE, Cardozo TJ, Zychlinsky A, Groisman EA. The selC-associated SHI-2 pathogenicity island of *Shigella flexneri*. Mol Microbiol 1999; 33:74–83.
51. Schmidt H, Zhang WL, Hemmrich U, et al. Identification and characterization of a novel genomic island integrated at selC in locus of enterocyte effacement-negative, Shiga toxin-producing *Escherichia coli*. Infect Immun 2001; 69:6863–6873.
52. Wieler LH, McDaniel TK, Whittam TS, Kaper JB. Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of the strains. FEMS Microbiol Lett 1997; 156:49–53.
53. Sperandio V, Kaper JB, Bortolini MR, Neves BC, Keller R, Trabulsi LR. Characterization of the locus of enterocyte effacement (LEE) in different enteropathogenic *Escherichia coli* (EPEC) and Shiga-toxin producing *Escherichia coli* (STEC) serotypes. FEMS Microbiol Lett 1998; 164:133–139.
54. Hensel M, Nikolaus T, Egelseer C. Molecular and functional analysis indicates a mosaic structure of *Salmonella* pathogenicity island 2. Mol Microbiol 1999; 31:489–498.
55. Bergthorsson U, Ochman H. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol Biol Evol 1998; 15:6–16.

56. Ochman H, Bergthorsson U. Rates and patterns of chromosome evolution in enteric bacteria. Curr Opin Microbiol 1998; 1:580–583.
57. Boyd EF, Hartl DL. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J Bacteriol 1998; 180:1159–1165.
58. LeClerc JE, Li B, Payne WL, Cebula TA. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science 1996; 274:1208–1211.
59. Makino K, Yokoyama K, Kubota Y, et al. Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak. Genes Genet Syst 1999; 74:227–239.
60. Mosig G, Yu S, Myung H, et al. A novel mechanism of virus-virus interactions: bacteriophage P2 Tin protein inhibits phage T4 DNA synthesis by poisoning the T4 single-stranded DNA binding protein, gp32. Virology 1997; 230:72–81.
61. Myung H, Calendar R. The *old* exonuclease of bacteriophage P2. J Bacteriol 1995; 177:497–501.
61a. Davis BM, Waldor MK. Filamentous phages linked to virulence of *Vibrio cholerae.* Curr Opin Micribiol 2003; 6:35–42.
62. Donohue-Rolfe A, Acheson DW, Keusch GT. Shiga toxin: purification, structure, and function. Rev Infect Dis 1991; 13(Suppl 4):S293–S297.
63. Plunkett G 3rd, Rose DJ, Durfee TJ, Blattner FR. Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. J Bacteriol 1999; 181:1767–1778.
64. Wagner PL, Neely MN, Zhang X, et al. Role for a phage promoter in Shiga toxin 2 expression from a pathogenic *Escherichia coli* strain. J Bacteriol 2001; 183:2081–2085.
65. Beutin L, Stroeher UH, Manning PA. Isolation of enterohemolysin (Ehly2) -associated sequences encoded on temperate phages of *Escherichia coli.* Gene 1993; 132:95–99.
66. Boyd EF, Brussow H. Common themes among bacteriophage - encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol 2002; 10:521–529.
67. Parkhill J, Wren BW, Thomson NR, et al. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 2001; 413:523–527.
68. Hansen-Wester I, Hensel M. *Salmonella* pathogenicity islands encoding type III secretion systems. Microbes Infect 2001; 3:549–559.
69. Lostroh CP, Lee CA. The *Salmonella* pathogenicity island-1 type III secretion system. Microbes Infect 2001; 3:1281–1291.
70. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella.* Mol Microbiol 2001; 39: 260–271.
71. Figueroa-Bossi N, Bossi L. Inducible prophages contribute to *Salmonella* virulence in mice. Mol Microbiol 1999; 33:167–176.
72. Figueroa-Bossi N, Coissac E, Netter P, Bossi L. Unsuspected prophage-like elements in *Salmonella typhimurium.* Mol Microbiol 1997; 25:161–173.
73. Mirold S, Rabsch W, Tschape H, Hardt WD. Transfer of the *Salmonella* type III effector *sopE* between unrelated phage families. J Mol Biol 2001; 312:7–16.
74. Hardt WD, Chen LM, Schuebel KE, Bustelo XR, Galan JE. *S. typhimurium* encodes an activator of Rho GTPases that induces membrane ruffling and nuclear responses in host cells. Cell 1998; 93:815–826.
75. Swartley JS, Marfin AA, Edupuganti S, et al. Capsule switching of *Neisseria meningitidis.* Proc Natl Acad Sci USA 1997; 94:271–276.
76. Dillard JP, Caimano M, Kelly T, Yother J. Capsules and cassettes : genetic organization of the

capsule locus of *Streptococcus pneumoniae*. Dev Biol Stand 1995; 85:261–265.
77. Dillard JP, Yother J. Genetic and molecular characterization of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 3. Mol Microbiol 1994; 12:959–972.
78. Townsend SM, Kramer NE, Edwards R, et al. *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. Infect Immun 2001; 69:2894–2901.
79. Tsolis RM, Townsend SM, Miao EA, et al. Identification of a putative *Salmonella enterica* serotype typhimurium host range factor with homology to IpaH and YopM by signature-tagged mutagenesis. Infect Immun 1999; 67:6385–6493.
80. Miao EA, Miller SI. A conserved amino acid sequence directing intracellular type III secretion by *Salmonella typhimurium*. Proc Natl Acad Sci USA 2000; 97:7539–7544.
81. Bakshi CS, Singh VP, Wood MW, Jones PW, Wallis TS, Galyov EE. Identification of SopE2, a *Salmonella* secreted protein which is highly homologous to SopE and involved in bacterial invasion of epithelial cells. J Bacteriol 2000; 182:2341–2344.
82. Wood MW, Jones MA, Watson PR, et al. The secreted effector protein of *Salmonella dublin*, SopA, is translocated into eukaryotic cells and influences the induction of enteritis. Cell Microbiol 2000; 2:293–303.
83. Zhang S, Santos RL, Tsolis RM, et al. The *Salmonella enterica* serotype typhimurium effector proteins SipA, SopA, SopB, SopD, and SopE2 act in concert to induce diarrhea in calves. Infect Immun 2002; 70:3843–3855.
84. Kingsley RA, Baumler AJ. Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. Mol Microbiol 2000; 36:1006–1014.
85. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001; 409:1007–1011.
86. Reeves MW, Evins GM, Heiba AA, Plikaytis BD, Farmer JJ 3rd. Clonal nature of *Salmonella typhi* and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori* comb nov. J Clin Microbiol 1989; 27:313–320.
86a. Kidgell C, Reichard U, Wain J, et al. *Salmonella typhi,* the causative agent of typhoid fever, is approximately 50,000 years old. Infect Genet Evol 2002; 2:39–45.
87. Andersson DI, Hughes D. Muller's ratchet decreases fitness of a DNA-based microbe. Proc Natl Acad Sci USA 1996; 93:906–907.
88. Perry RD, Fetherston JD. *Yersinia pestis*—etiologic agent of plague. Clin Microbiol Rev 1997; 10:35–66.
89. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc Natl Acad Sci USA 1999; 96:14,043–14,048.
90. Sanderson KE, Hessel A, Liu S, Rudd KE. The genetic map of *Salmonella typhimurium*, edition VIII. In: Neidhardt FC, Curtiss R (eds). *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. 2nd ed. Washington, DC: ASM Press, 1996, pp. 1903–1999.
91. Berlyn MKB, Brooks Low K, Rudd KE. Linkage map of *Escherichia coli* K12, Edition 9. In: Neidhardt FC, Curtiss R (eds). *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. 2nd ed. Washington, DC: ASM Press, 1996, pp. 1715–1902.
92. Tillier ER, Collins RA. Genome rearrangement by replication-directed translocation. Nat Genet 2000; 26:195–197.
93. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 2000; 1:RESEARCH0011.
94. Deng W, Burland V, Plunkett G 3rd, et al. Genome sequence of *Yersinia pestis* KIM. J Bacteriol 2002; 184:4601–4611.

95. Roth JR, Benson N, Galitski T, Haack K, Lawrence JG; Miesel L. Rearrangements of the bacterial chromosome: Formation and applications. In: Neidhardt FC, Curtiss R (eds). *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. 2nd ed. Washington, DC: ASM Press, 1996, pp. 2256–2276.
96. Mackiewicz P, Mackiewicz D, Gierlik A, et al. The differential killing of genes by inversions in prokaryotic genomes. J Mol Evol 2001; 53:615–621.
97. Tamas I, Klasson L, Canback B, et al. Fifty million years of genomic stasis in endosymbiotic bacteria. Science 2002; 296:2376–2379.
98. van Ham RC, Kamerbeek J, Palacios C, et al. Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci USA 2003; 100:581–586.
99. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. Nature 2000; 407:81–86.
100. Akman L, Yamashita A, Watanabe H, et al. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat Genet 2002; 32:402–407.

17
专性细胞内病原体

Siv G. E. Andersson

引言

真核细胞是一个可以为细菌提供养分和保护，并具有非常吸引力的生长栖息地。然而，为了能够在另一种生物内生存和繁殖，入侵细菌必须进入寄主，并在一些特定的细胞中增殖，最终逃逸并在另一寄主中重复其生活史[1]，在这个过程的所有阶段，细菌必须避免被寄主的免疫系统杀伤[1]，因此，细菌向细胞内生长环境的迁移不是一个简单的过程。专性细胞内寄生物（obligate intracellular parasite）定义为以严格寄主关联方式增殖的细胞内细菌[1]，而兼性细胞内寄生物（facultative intracellular parasite）虽然能在细胞内生存，但仍保留了在寄主体外生长的能力[1]。兼性细胞内寄生和专性细胞内寄生的区别在于这些生物是否能在体外培养[1]，这种生活方式的差异至今还没有在分子水平上得到解释。

可想而知，专性寄生细菌缺乏自由生活能力的原因，在于缺乏编码某些关键酶的基因或缺乏由细胞内提供的某些化合物，实际上，很多这种依赖性已被证实。尽管已经对专性细胞内细菌在人工培养基中的生长做了大量尝试，但仍然无法在实验室的体外条件下培养；另一原因可能是寄生菌的复制和细胞分裂受寄主细胞的信号调节，但是目前还没有找到这种潜在的调控分子。不论对专性寄主依赖性的基础如何，真核细胞内部依然是细菌入侵的最具吸引力的目标，许多在进化上互不相关的种属，已经各自独立地探索过这个生长栖息地，其实这并不值得惊奇，真核细胞内有大量代谢产物，也几乎或完全不存在与其他细菌的竞争，然而，进入这种优越环境所付出的代价直接在核苷酸水平上表现出来，在漫长的时期内，导致完全依赖寄主和严重的基因组退化[2]。

立克次氏体和衣原体

衣原体（*Chlamydiae*）和立克次氏体（*Rickettsiae*）两个模式系统对理解专性细胞内寄生行为非常重要。过去，立克次氏体用来命名许多微小而又不能体外培养的棒状细菌[3]，通过对某些特定基因的分析，如核糖体核酸（rRNA）基因的分子序列，发现它是许多不同种群细菌的组合体，包括立克次氏体、考克斯氏体（*Coxiella*）和埃里希氏体（*Ehrlichia*）[4]。20 世纪 40 年代，建立了用鸡胚卵黄囊培养立克次氏体的方法，便能制备用于精确生化研究的足够材料。尽管有了这些进展，生物学、流行病学和种系发生学的进展仍然缓慢，主要原因是这些细菌的苛刻生长需求和高致病性。在 20 世纪 90 年代初，每种生物中只有几十个基因在分子水平上进行了特征分析。

到 20 世纪 90 年代中期，DNA 自动测序仪提供了细菌、古生菌和真菌的基因组数

据，为微生物学带来了一场革命。两种专性细胞内寄生物，普氏立克次氏体和砂眼立克次氏体的全基因组序列首次在 1998 年发表[5,6]，与之相关的几个菌株和种群在基因组水平进行了特征分析[7~10]，遗传操作也于 1998 年取得成功，立福平霉素抗性基因转入到普氏立克次氏体基因组[11]，这为立克次氏体遗传系统的发展迈出了第一步。

未来的挑战是如何利用和开发基因组数据和提高其转化能力，以深入了解立克次氏体的体内行为特征。本章的目的是介绍如何利用所获得立克次氏体全基因组序列，促进对它们的生理和进化方面的研究，并着重介绍普氏立克次氏体和康氏立克次氏体的基因组。不仅如此，还将更广泛地讨论专性细胞内寄生行为的普遍性概念，包括对其他专性细胞内寄生物的讨论。

流行病学、种系发生学和病症

斑疹伤寒病原的拉丁学名 *Rickettsia prowazekii*（普氏立克次氏体）是为了纪念两位首次发现这种病原的微生物学家，H. T. Ricketts[12]和 S. J. M. Prowazek。不幸的是，这两位科学家都患了斑疹伤寒，在试图揭示斑疹伤寒病原奥秘的过程中去世。在全球范围内，斑疹伤寒病原在几百年的时间里都是人类的灾难，它导致上千万人死亡。流行性斑疹伤寒临床症状的最早描述出现在 16 世纪地中海地区，在那里，疾病一直跟随着穿越欧洲的军队，接着通过受感染的虱子传染给平民，这都是 Nicolle 在 1909 年发现的[13,14]，关于斑疹伤寒通过人体上虱子（*Pediculus humanus corporis*）传染的发现为他赢得了 1928 年诺贝尔奖。这一发现在第一和第二次世界大战中发挥了最大的作用，当时，刮胡子、洗涤和焚烧衣物是将传染病降低到最小的重要卫生措施。尽管现在流行性斑疹伤寒的爆发很少，但是，根据世界卫生组织的报告，这种疾病仍然是一些非洲国家的主要问题，最近一次是 1995 年爆发的，由于布隆迪的卫生条件差，暴发了由虱子传染的斑疹伤寒，当时这种疫病在该国的大部分地区蔓延[15]。

人、媒介和储主

人是主要的寄主，也是目前所知普氏立克次氏体的唯一天然储主，普氏立克次氏体利用人体上的虱子作为传播的媒介[16]。虱子是严格的吸血昆虫（图 1），有上百万个细菌细胞被分泌到虱子叮咬的皮肤伤口周围，通过摩擦或者抓挠这些细菌就进入伤口引发致命的疾病。令人惊异的是，病原菌的感染对传播疾病的虱子也是致命的，这是因为普氏立克次氏体能够在虱子的中肠上皮细胞内快速繁殖，引发细胞的裂解。根据肠道中细菌的数量，虱子会在 1～2 周内死亡，而正常虱子的生命周期为 1 个月。

大多数立克次氏体的生活史都包括动物寄主和跳蚤、螨、扁虱等节肢动物媒介（表 1）[16,17]，仅有几种立克次氏体已经与它们的节肢动物媒介建立了共生关系，由卵传播的方式代代相传（即由感染雌虫产受感染的卵）[18]。尽管一个跳蚤在实验室能够同时被两种斑疹伤寒立克次氏体（*Rickettsia typhi*）和猫立克次氏体（*Rickettsia felic*）感染，但是在自然条件下还未被发现两种或多种立克次氏体的共感染。感染普氏立克次氏体的扁虱能够抵抗立氏立克次氏体（*Rickettsia rickettsii*）的感染，表明立克次氏体能诱发细胞质内的不相容性，然而，还没有种系发生学的证据来支持扁虱和立克次氏体的长期相

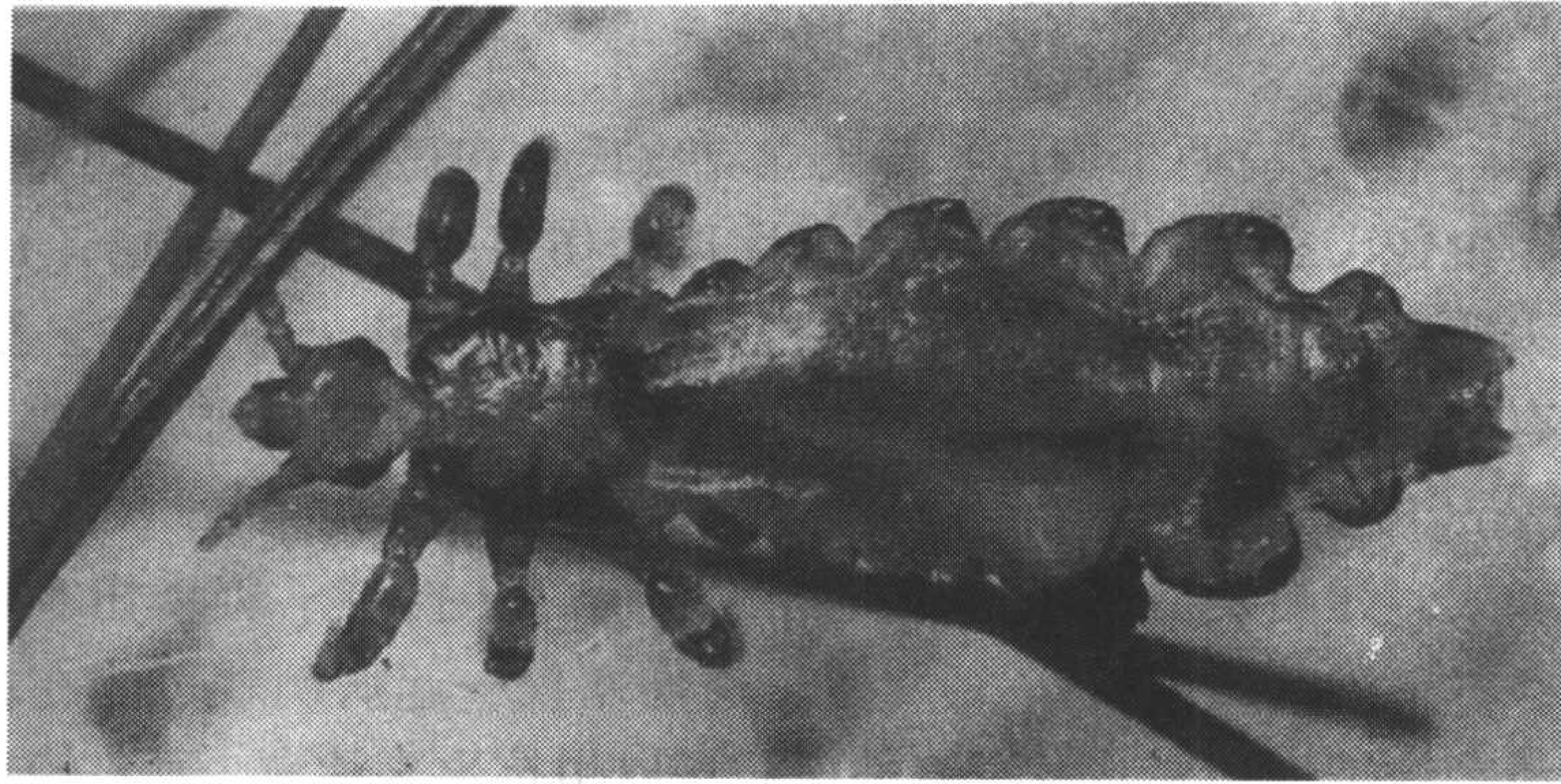


图 1　人体虱子（*Pediculus humanus*）——专性细胞内病原普氏立克次氏体的寄主和载体。

表 1　立克次氏体种类、载体和人类疾病

群	种	疾病	载体
TG	普氏立克次氏体(*R. prowazekii*)	流行性斑疹伤寒	人体虱
TG	斑疹伤寒立克次氏体(*R. typhi*)	鼠型斑疹伤寒	鼠跳蚤
SFG	康氏立克次氏体(*R. conorii*)	地中海斑疹热	扁虱
SFG	西伯利亚立克次氏体(*R. sibirica*)	西伯利亚扁虱伤寒	扁虱
SFG	立氏立克次氏体(*R. rickettsii*)	洛基山斑疹伤寒	扁虱
SFG	蒙大拿立克次氏体(*R. montana*)	未知	扁虱
SFG	扁头蜱立克次氏体(*R. rhipicephali*)	未知	扁虱
SFG	澳大利亚立克次氏体(*R. australis*)	鹦鹉热	扁虱
SFG	螨立克次氏体(*R. akari*)	立克次氏体痘	螨
——	加拿大立克次氏体(*R. canada*)	未知	扁虱
——	贝利立克次氏体(*R. bellii*)	未知	扁虱
——	AB 杆菌(AB bacterium)	未知	甲虫

注：TG，斑疹伤寒群；SFG，斑疹热群；-，未分类。

互作用和共进化[19]，然而，已经发现了蚜虫与它们共生体的这种关系[20]。

立克次氏体种系发生学背景

由 rRNA 基因的 DNA 序列推断，立克次氏体种系发生学位于 α 多形菌[4]，它主要由两个群组成：斑疹伤寒群（typhus group，TG）立克次氏体和斑疹热群（spotted fever group，SFG）立克次氏体（表 1）[19]。有几个种，如贝利立克次氏体（*Rickettsia bellii*）和加拿大立克次氏体（*Rickettsia canada*）在种系发生学上与这两个主群接近，但是又有显著区别[21,22]。某些与这两个主群差别更大的种最近已将它们从恙虫热立克次氏体（*Rickettsia tsutsugamushi*）重新归为恙虫病东方体（*Orientia tsutsugamushi*），这强调了它们与其他立克次氏体有更早的分化[23,24]。

斑疹伤寒群立克次氏体

TG 立克次氏体群只有两个成员：普氏立克次氏体和斑疹伤寒立克次氏体，它们对人都有致病性。由普氏立克次氏体引发流行性斑疹伤寒的潜伏期为 10～14 天，其临床症状是高热和头痛，5～7 天后，躯干、四肢和腋下出现皮疹，中枢神经系统受到侵袭，神经性紊乱时有发生，患者陷入昏迷，在此期间伴随有体温高、血压低等症状，该病的致死率为 10%～30%。

以慢性病形式，如复发型斑疹伤寒（*Brill-Zinsser*）携带了斑疹伤寒病原，患者在有压力的生活条件下激活病症[19]，复发型斑疹伤寒仅一起病例就可能在虱子横行的人群中引发新的流行性斑疹伤寒。幸亏这种疾病可以用抗生素治愈，而在自然条件下立克次氏体的抗药性还未发现。由于流行病的爆发严格依赖虱子的出没，改善卫生措施在上个世纪已大大控制了疾病的蔓延。

人群中温和形式的斑疹伤寒-鼠斑疹伤寒-是由斑疹伤寒立克次氏体引起的，这是 TG 群的另一个成员，大鼠是该病原的主要储主，通过鼠或鼠虱传染人。根据综述的目的，本文将普氏立克次氏体作为 TG 立克次氏体的代表，因为它的完整基因组序列已被测定[5]。

斑疹热群立克次氏体

目前 SFG 立克次氏体的成员超过 20 种，包括康氏立克次氏体（*Rickettsia conorii*），它是地中海斑疹热的病原体，这种专性细胞内病原通过狗褐扁虱传给人，病征表现为高热、头痛、肌痛和关节痛。另一种人类病原是立氏立克次氏体，它是引发洛基山斑疹热的病因，与康氏立克次氏体一样，立氏立克次氏体也是通过受感染扁虱的叮咬传给人，这种病引起发烧和全身出疹[25]，它通常在北美、南美和中美发生，不治疗将导致死亡。

SFG 成员引起的其他疾病有非洲扁虱斑疹伤寒和立克次氏体痘疹。有趣的是致病种，如立氏立克次氏体、帕氏立克次氏体（*Rickettsia parkeri*）和西伯利亚立克次氏体（*Rickettsia sibirica*）在种系发生上与非致病种，如扇头蜱立克次体（*Rickettsia rhipicephali*）和蒙大拿立克次氏体（*Rickettsia montana*）不同，其他种如猫立克次氏体和 *Richettsia helvetica* 在 SFG 群中早成为另一分支。致病种与非致病种在分子水平上的差异还没有彻底阐明，该群中的康氏立克次氏体的全基因组序列数据已经齐备，本综述把康氏立克次氏体作为 SFG 群的代表种[7]。

吸附、侵入、增殖和释放

立克次氏体通过诱导性吞噬作用进入寄主细胞，它与寄主细胞的受体相结合形成吞噬囊泡。尽管立克次氏体能够在体外条件下进入不同的有核细胞，但是在体内它的主要目标是上皮细胞，通过依赖肌动蛋白过程侵入[26]，一旦进入细胞，细菌就会先于吞噬体-溶酶体的融合（phagosome-lysosome fusion）而诱发吞噬体膜（phagosomal membrane）的裂解，从而进入细胞质。立克次氏体在所有专性细胞内病原中是独一无二的，它可以直接在寄主的细胞质内增殖而不需要被任何由寄主产生的膜所包被。某些种，如加拿大立克次氏体甚至能够在细胞核内增殖，在增殖过程中，病原只会给寄主细胞造成

中度损害。细胞死亡的根本原因在于，寄主细胞质的空间不够容纳内部过多寄生病菌而造成的物理性裂解。

某些种，如康氏立克次氏体和立氏立克次氏体能够将肌动蛋白聚合成慧尾结构[27,28]，已观察到它们能在受感染的细胞质中运动，这与李斯特氏菌（*Listeria*）和志贺氏菌（*Shigella*）基于肌动蛋白的运动相似。这个过程有助于促进由起始侵染位点开始的细胞间传播，而且不需要任何进一步的胞外阶段[29]。已经发现立克次氏体诱导内皮细胞培养物的空斑，这说明细胞间的传播伴随有嗜细胞作用。然而，上皮细胞的感染能够诱发寄主细胞存活至关重要的抗衰老作用（antiapoptotic effects)。看来立克次氏体能够调整寄主细胞的过敏毒素反应，保留寄主细胞作为侵染位点为其服务[30]。

立克次氏体基因组的结构

普氏立克次氏体和康氏立克次氏体的基因组都很小，分别为 1.11Mb 和 1.27Mb（表 2）[5,7]，两个基因组的整体结构本质上相同，只是在复制终止区附近有几个重排（图 2）。在衣原体中发现了位于复制子和终止子周围区域的脱氧核糖核酸对称倒转序列[31]，这些重排的对称特征是开放复制叉上发生重组的结果。这种移位和倒置在许多其他基因组中发现，这说明开放复制叉部位的 DNA 复制时很容易发生重组。

表 2　普氏立克次氏体[5]和康氏立克次氏体[7]基因组特征的比较

特征	普氏立克次氏体	康氏立克次氏体
基因组大小/bp	1 111 523	1 268 755
G+C 遗传含量/%	29.0	32.4
编码蛋白基因/个	834	1374
假基因/个	>12	>2
非编码含量/%	<76	<81

这两种立克次氏体基因组大小的差异，表现出它们基因含量的不同：普氏立克次氏体有 834 个明确的基因[5]，康氏立克次氏体有 1372 个[7]，康氏立克次氏体基因组中有普氏基因组中 834 个基因中的 803 个，另外 552 个可读框为康氏立克次氏体独有[7]。但是，后来对普氏立克次氏体基因组的仔细研究发现，在康氏立克次氏体那 552 个特有可读框中，有 229 个基因为残余序列。由此可推断，普氏立克次氏体从康氏立克次氏体中分化出去后，已彻底删除了 200 多个基因，另外 200 多个基因也严重退化。

自普氏立克次氏体从康氏立克次氏体中分化出来后，有近三分之一基因的编码能力丢失，从人的角度出发，非常值得注意具有高致病性的普氏立克次氏体，实质上是与之相近低致病性康氏立克次氏体的退化版本。

营养依赖、代谢和调控

与自由生活细菌所处的自然生长环境相比，真核生物细胞质的营养极为丰富，专性

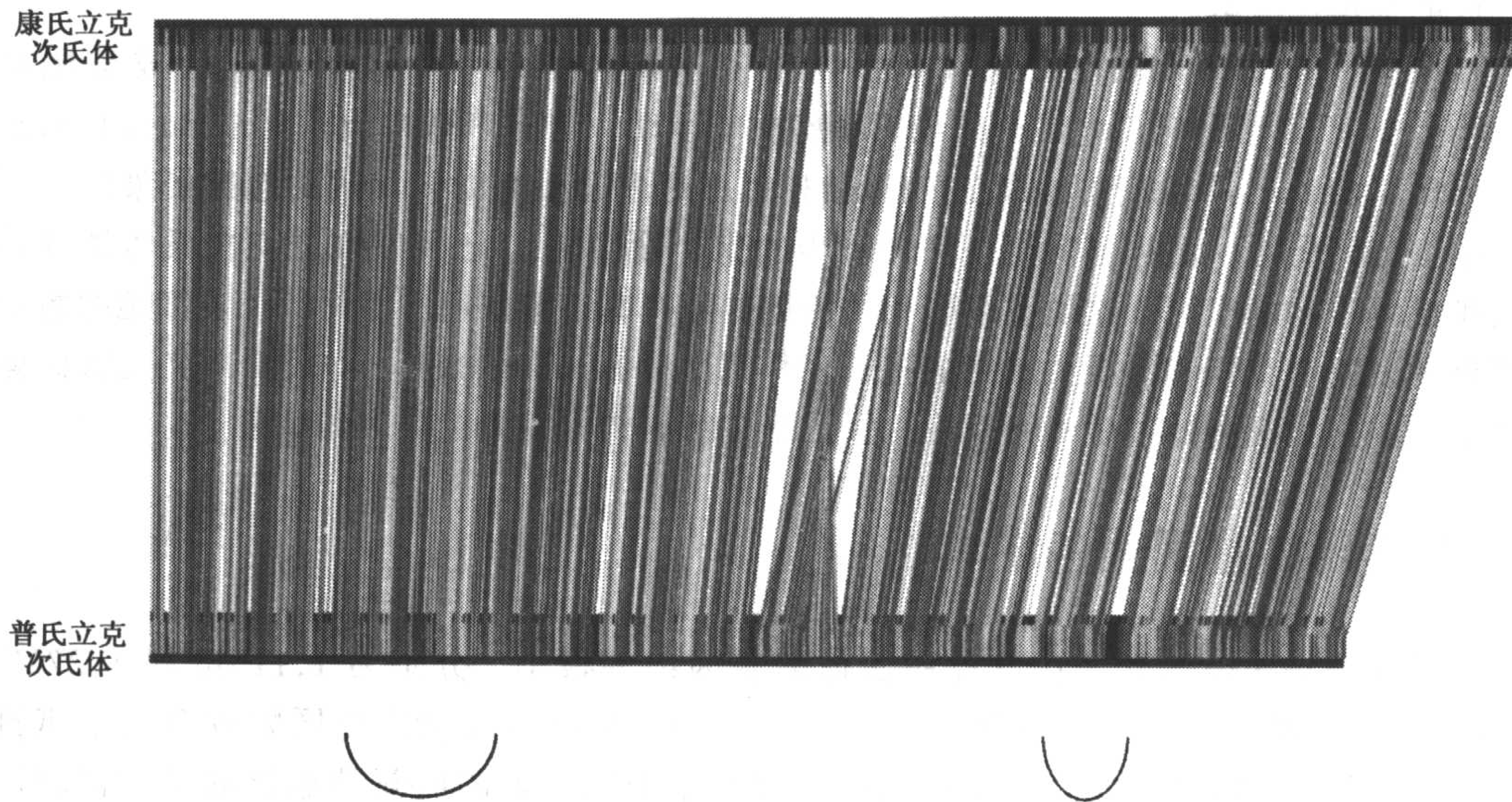


图 2　普氏立克次氏体和康氏立克次氏体的基因组示意图。两个基因组间的线表示基因同线性（synteny）区域，中部交叉线表示复制终止区附近的染色体重排，弓形线显示大于 100 个核苷酸重复序列的分布。

细胞内寄生微生物的生活方式与其近缘自由生活的细菌不同，它们只能在单一的生长环境——真核寄主细胞内繁殖，与之相反，自由生活细菌在营养过度和营养匮乏的条件之间不断切换，因此它们必须配备一系列生存措施，包括能够感知外部环境的改变并能做出相应反应的调控基因。专性细胞内病原与之不同，它只需要对少量的外部信号做出反应。的确，适应细胞内生活方式明显导致与调控过程相关大量基因的缺失，立克次氏体中的基因调控已经阐明，例如，几个关于三磷酸腺苷（ATP）生物合成和转运的基因、柠檬酸合成酶与 ATP/ADP 转运酶都受到 ATP 供应的调控。有关这些内容的更多细节将在另一部分讨论。

小分子生物合成基因的缺失

真核生物细胞质提供了诸如氨基酸和单磷酸核苷酸等各种小分子化合物，这就不奇怪为什么立克次氏体中缺乏编码合成氨基酸和核苷酸的基因[5,7]，正如所料，确认了立克次氏体中单磷酸核苷酸的转运系统，这一切就能得到解释在立克次氏体种群中另一个失去功能的生物合成途径是合成 S 腺苷甲硫氨酸（S-adenosylmethionine，SAM）[32]。同样，在普氏立克次氏体中也发现了 SAM 转运系统[32a]，立克次氏体中也没有糖酵解基因，从而推测丙酮酸来自真核生物细胞质，这些生物合成基因的缺失表明，立克次氏体完全依赖寄主提供各种小分子。

细胞质 ATP 的利用

专性细胞内病原最有趣的特征是它们从细胞质内运输 ATP 的能力，这是专性细胞

内寄生物，如细菌类的立克次氏体和衣原体[5~10]以及真核生物微孢子虫（*Encephalitozoon cuniculi*[33]）所独有的。与之类似的ATP/ADP转运系统也在质体（plastids）内膜发现[34~38]，线粒体也有能够介导ATP/ADP转运功能的膜蛋白，但是这种蛋白是从某种更普遍的膜蛋白类型演化来的，在种系发生学上不同于专性胞内寄生生物和质体[39]。没有一种自由生活细菌有这种转运ATP/ADP的功能，这不以为怪，因为转运的方向依赖于跨膜ATP/ADP的浓度。除了真核细胞内部，没有天然的生长环境富含ATP，这就意味着如果自由生活细菌有这样的膜转运蛋白，那么ATP就会排到外部环境中。

普氏立克次氏体和康氏立克次氏体的基因组含有多达5个ATP/ADP转运蛋白基因[5,7]，第一个基因*tlc1*，实验证实该基因用于催化ATP和ADP的转运[40]，普氏立克次氏体中基因*tlc1*的转录水平比寄主细胞质中的ATP水平敏感[40]，在细胞质中富含ATP的早期阶段其表达水平最高，但在被普氏立克次氏体严重感染的细胞中下降。这种表达模式与生产ATP基因（如柠檬酸合成酶基因）的表达模式不同，后者在寄主细胞中ATP减少的情况下大量表达[40]。因此，ATP/ADP转运酶的早期合成能使普氏立克次氏体有效地利用寄主细胞的ATP，而随后表达水平的下降可以避免其内部有氧呼吸产生的ATP渗漏到细胞质中。

在砂眼衣原体中也发现了编码核苷转运蛋白的并系同源蛋白——Npt1和Npt2的基因，这两个蛋白质在氨基酸水平上分别与普氏立克次氏体的ATP/ADP转运蛋白有68%和61%的相似性。Npt1以交换模式催化ATP和ADP的转运为砂眼衣原体供能，Npt2催化合成代谢所需三磷酸腺苷的净吸收（net uptake）[41]。从进化角度看，立克次氏体、衣原体和微孢子虫是由不同的微生物祖先演化来的，尽管如此，都具有ATP/ADP转运蛋白，也许是它们成功建立这种寄生关系的主要原因。

立克次氏体基因组的退化

普氏立克次氏体基因组中，有多达24%的基因组由非编码DNA序列组成[5]，这是其突出特点，注释后的康氏立克次氏体中非编码DNA估计占19%[7]，其他胞内专性寄生物，如麻风分枝杆菌（*Mycobacterium leprae*）含有更高的非编码区，约为56%[42]。正如开始所推测的那样[5]，普氏立克次氏体中的基因间隔序列由不具活性的退化基因组成，而这些序列还没有被完全删除[32,43,44]，例如，通过与其近缘菌株和菌种的比较分析发现这些区域中编码基因的缺失。从这种比较还发现某个种的基因与其他种的假基因非编码DNA相对应[32,43,44]。对普氏立克次氏体和康氏立克次氏体基因组的分析发现，康氏立克次氏体的552个可读框中所特有的229个在康氏立克次氏体中有残体(remnants)[7]，许多这些残体与其他种的基因没有序列同源性，成为孤体（orphan)。在大多情况下，孤体比其他有同源基因的残体短，表明一些孤体只能代表基因片段，而不能代表完整基因。

基因组上冗余DNA的获得及其进化在立克次氏体中研究最深入，这得益于这些基因组中的高A+T含量。在TG所有成员中，G+C含量为29%～30%，在SFG成员中其含量为32%～33%。突变多发生在A+T含量高的非编码区和密码子的第三个核苷酸，该位置的G+C含量可低至10%～15%。密码子使用频率与预测的表达水平之间没

有相关性，表明密码子使用频率的自然选择在这些种群中不是很有效[45]。编码区核苷酸频率缺乏统计学多样性，这极大地简化了对任何特定序列编码状态的判断，编码重要基因功能的序列比非编码区有更高含量的 G+C 和低替换频率。下面将讨论由基因组比较而推断出的基因组片段和基因序列从立克次氏体基因组中被逐渐剔除的过程。

多拷贝基因序列的缺失

由于冗余基因的缺失不见得有致死效应，因此，多拷贝基因是细菌适应胞内生长环境过程中首先被删除的对象。确实，胞内专性寄生物的小基因组含有极少重复序列，没有或只有很少遗传寄生序列（genetic parasite）[46]。与之相反，自由生活细菌的大染色体组拥有大量诸如转座子和细菌噬菌体的重复序列和自我增殖的 DNA 序列[46]。多个基因组的总体分析解释了基因组大小、重复序列和生活方式之间的相关性，专性寄生细菌，如立克次氏体、衣原体和巴克纳氏菌便是一些极端的例子[47]。

剔除冗余基因首先是剔除编码 rRNA（*rrs*、*rrl* 和 *rrf*）和延长因子 Tu（*tuf*）基因，这些基因在一般细菌基因组中有 2 个或更多拷贝数[48]。但是，在立克次氏体中只有一个 rRNA 和 *tuf* 基因[49~51]，对几种立克次氏体的比较研究发现，在这些立克次氏体属分化前便发生 rRNA 和 tuf 副本的丢失[52]。基因 rRNA 和 *tuf* 的副本不仅是序列删除的目标，同时也是基因组重排的目标，实际上，据推测立克次氏体中一个 *tuf* 基因的倒置是由两个祖先 *tuf* 基因在染色体内的重组造成的，随后又将其中一个拷贝删除[50]。多拷贝基因组的重组不仅导致了重排，而且造成立克次氏体古老基因组在进化早期大片段 DNA 的删除，这使立克次氏体在适应细胞内生活的早期就快速地删除了序列。

短重复序列的缺失

两种立克次氏体基因组之间最显著的区别是康氏立克次氏体基因组比普氏立克次氏体多了很多短重复序列（图 2）[7]，在康氏立克次氏体基因组中发现了 10 个重复家族的总共 656 个重复序列，它们的大小在 19～172bp 之间，占整个基因组 3.2%，这些重复序列富含 G+C（约 40%），从而造成这两个基因组间 G+C 含量的差异，普氏立克次氏体为 29%，而康氏立克次氏体为 32%（表 2）。

通过立克次氏体基因间区域的比较研究发现，一个或几个种的基因组上的短重复序列，经常出现在另一个种中被删除的短序列两侧[32,44,51,52]，如果这些删除是由短重复序列介导，那么在这些含有删除片段的菌株中只有一个拷贝的重复序列，这些重复序列的大小从 7～30bp 不等，分布在整个基因间区域，例如立氏立克次氏体一个约 7bp 短重复序列位于 *fmt-rrl* 5′端的间隔区域，连接该种特有的 28bp 序列[51]，同样的，一段 7bp 短重复序列紧挨着分别在蒙大拿立克次氏体和猫立克次氏体中缺失的一段序列[52]。这说明短重复序列扮演着一个重组目标的角色，它不但引起目标片段的缺失，而且还导致一个拷贝重复序列的缺失。普氏立克次氏体的基因组比康氏立克次氏体小，并且含有更少的重复序列，据此推断，由于重组频率高，普氏立克次氏体的短序列成分比康氏立克次氏体消耗得快。

立克次氏体中倒转重复序列的退化

特别令人感兴趣的一种重复序列类型是立克次氏体的回文序列（palindromic ele-

ment, RPE), 据报道, 康氏立克次氏体基因组中有 45 个拷贝, 其中 19 个定位于基因内部[53,54], 因为重复序列一般在非编码区, 而这些出现在基因内部的重复序列格外引人注目。当用康氏立克次氏体的 RPE 在普氏立克次氏体中寻找类似序列时, 在普氏立克次氏体中也发现了 10 个在基因内部高度分化的 RPE[53]。更加深入全面地在康氏立克次氏体基因组[7]中寻找重复序列时, 发现了总共 656 个散布的重复序列, 它们被分为 10 个不同的家族[53]。在康氏立克次氏体基因组中, 许多这些重复序列的全长或一部分位于那些被注释为基因的 ORF 中[55]。

RPE 最早被看作是基因组内部增殖的产物[53,54]。然而, 在许多立克次氏体种属内, 包括 TG 和 SFG 中的成员都发现了相同的 RPE 插入位点, 这说明 RPE 是在立克次氏体属内分化之前获得的[52,55], 很多 RPE 都不完整。对立克次氏体多个种间的 RPE 比较推测, 这些序列在几个种中已经退化[52], 如果这样, 那就说明某些古老的 RPE 已经缺失, 或者不能被序列相似性查询所发现。实际上, 普氏立克次氏体和斑疹伤寒立克次氏体中高度残缺的 RPE, 是通过对其他种 RPE 插入位点的同源性定位所证实[52], 因此, RPE 作为目标寄主蛋白基因组的特异性, 最简单的解释是在一些种内缺失, 而不是像其他种内的增殖[52,55]。

康氏立克次氏体的 RPE 被认为是通过插入 RNA 和编码蛋白的基因而对蛋白进化做出贡献[53,54], ORF 内的插入位点能被蛋白的三维结构和功能所容忍, 有两种类型的 RPE 被预测能产生 α 螺旋和 β 折叠[55]。然而, 也不能排除这些结构的形成受插入位点周围环境对结构影响的可能性[55]。确实, RPE 在立克次氏体属内缺乏保守性, 说明 RPE 是一种中性基因组乘客, 可能不对任何功能产生影响[52]。

如果所有发现康氏立克次氏体的 RPE 起源于立克次氏体的属内分支, 那么, RPE 在 TG 中消失肯定比在 SFG 中快。这便联想到在普氏立克次氏体中基因的缺失比在康氏立克次氏体中更严重[7], 康氏立克次氏体中总共有 500 多个的独特基因[7], 在普氏立克次氏体基因组中发现了其中约有 200 个对应基因的残体[5], 这表明基因缺失在普氏立克次氏体中进行得更快。因此, 与普氏立克次氏体相比, 尽管康氏立克次氏体基因组更大, 含有更多重复序列, 但它仍然必须遵循与普氏立克次氏体相同的退化规律。然而, 在这两种立克次氏体中, 删除突变的固定频率 (fixation rate) 不同, 这可能是由代时或种群结构不同所造成。

单基因序列的退化

从编码 S 腺苷甲硫氨酸 (S-adenosylmethionine, SAM) 合成酶的 *metK* 基因中, 获得了第一个关于基因序列退化细节的信息, 该酶是生物合成 SAM 辅助因子所必需的[32]。在普氏立克次氏体菌株 Madrid E 中, 这个基因含有一个终止密码子, 而在普氏立克次氏体菌株 Breinl 和斑疹伤寒立克次氏体中的可读框却是开放的。在其他所有立克次氏体中, 这个基因是以随机方式累积突变。在立克次氏体中发现的 SAM 转运系统, 可能导致基因 *metK* 成为非必需基因[32a]。

基因 *metK* 的突变似乎是中性突变, 主要是删除了一个或几个碱基[32]。在 26 个失活基因的研究中发现, 多达 1536 个核苷酸被删除掉, 只有 31 个核苷酸被插入[44], 大多数被删除序列都很短, 多数核苷酸的缺失是由两个 599bp 和 767bp 大片段的删除而引

起的。在立克次氏体的每个突变中，删除平均值和中值预测分别为 4 个和 51 个核苷酸[43,44]。对删除的强烈嗜好表明，通过累积内部终止子和移码突变而造成的失活基因最终被切除，而切除的速度是复制-修复机制产生突变频率的函数。

弱残基因的表达

康氏立克次氏体基因组中的一小部分基因是近期才发生突变的，就像普氏立克次氏体的基因 *metK* 一样[7]。在康氏立克次氏体中发现了一些短 ORF，它们与其他种的一些全长直系同源基因（ortholog）相似，包括普氏立克次氏体中的某些基因。总之，37 个古老基因已经退化成多达 105 个短 ORF[7]。在这些基因中，在普氏立克次氏体中有 14 个完整的直系同源基因，其余 23 个在普氏立克次氏体中没有发现。反之亦然，在普氏立克次氏体基因组中有 11 个基因被割裂成 23 个 ORF，所有这些 ORF 在康氏立克次氏体中都有完整的直系同源基因[7]。

奇怪的是，有几个（不是全部）这样的 ORF 能够合成 RNA，这表明某些这样的短 ORF 仍具有功能[7]，为了研究康氏立克次氏体中这些基因片段的功能，便分析了这些片段的替代频率[56]，迄今为止，已经比较了这些断裂基因对全长基因的替换频率、不同表达特点断裂基因的替换频率，以及不表达和表达基因片段的替换频率。据统计，直系同源全长基因的非同义替换总频率（代表替换位点处氨基酸的平均替换值）为 0.07[56]，无论 ORF 表达与否，断裂基因片段的替换频率都较高，这表明这些短 ORF 不具有功能[56]。

由此推断，这些 ORF 的启动子来源于突变或断裂基因中存在的序列，由于立克次氏体基因组中仅含很少调控基因，因而对转录的控制不严谨，难以阻止起始 A/T 富含区的无意义转录，尤其是退化基因的转录。确实，细菌启动子富含 A/T，潜在启动子序列在富含 A/T 的普氏立克次氏体和康氏立克次氏体基因组中很常见[5,7]。从原理上推断，这可能导致残留基因有部分功能，从而弥补了终止密码子的引入和移码造成的突变，同时，转录也能被随机启动，但在蛋白质水平上没有任何功能。

可以认为，康氏立克次氏体中的断裂基因，代表已经开始累积突变的退化基因[56]，在非同义位点替换频率的增加表明，断裂基因不再具有功能，其中某些片段的表达很可能只是一种暂时现象[56]，是否少数被表达的断裂基因仍具有某种功能还有待进一步研究。

这些基因组最终会灭绝吗?

迄今为止，对所有序列数据，包括基因、假基因、非编码 DNA 和重复序列的分析表明，TG 中序列删除的过程进行得比 SFG 中的快。此前对以中性方式突变的序列研究表明，在两种立克次氏体中都存在删除突变的倾向，并且涉及的片段大小比插入突变大[32,41]。然而，在 SFG 中进化的基因，其存活的可能性比在 TG 中进化的基因要大，由于这种差异以至在 TG 和 SPG 中的失活 DNA 中也可以观察到，因此，这种差异可能是由删除频率的不同所造成，而不是基因产物的功能不同所造成。

TG 和 SFG 立克次氏体之间的另一个主要差异，是在 TG 中没有发现高比例 SFG 中的短重复序列和反向重复序列，如上所讨论的[48,50,51]，短重复序列作为重组的热点，

导致相关序列的倒转或删除，甚至少于 10 个核苷酸非常短的重复序列也能介导这一过程[51,52]。即使重组介导删除发生的频率比每次删除几个核苷酸的突变频率低得多，由于重复序列的重组影响到大量核苷酸的删除，成为导致两种立克次氏体间差异的主要原因。

因此，基因组大小和退化频率的差异，可能与某种导致删除突变的内在机制有关，或与不同种群结构的差异有关，例如，传播康氏立克次氏体的扁虱和传播普氏立克次氏体的虱子有显著差异，扁虱的寿命是 5～6 年，每年交配繁殖一次，而在它的一生中只吸血两次；而虱子的一生只一个月，每天交配繁殖一次，每天吸血五次，因而菌群传播频率虱子比扁虱高。如果切除突变主要在寄生生活史的复制期产生，利用短生活周期媒介的病原每年会有更多世代，这就提高了在种群中删除突变类型的可能性。

尽管如此，在这两种情况下，最初基因组减少似乎是发生在重复序列的同源重组上，其结果是，DNA 大片段的消除与长重复序列的删除相伴随进行，接下来越来越短重复序列在重组过程中也逐渐被消耗，直到像现在的普氏立克次氏体那样，序列的消除主要是由遗传序列中短删除的累积造成的，由每次只几个核苷酸的短删除，造成单基因失活频率较低，导致基因组中假基因和冗余基因的暂时增加。

总之，迄今获得的数据表明，康氏立克次氏体和普氏立克次氏体面临着相似的进化压力，但是它们生活史的不同造成这些过程以不同的频率进行，这表明康氏立克次氏体基因组正在缓慢地演化成现在的普氏立克次氏体，而普氏立克次氏体基因组到那时还会变得更小，最终，所有重复序列被耗尽，基因的退化将会以极其微小的步伐进行。

这样的推测意味着基因组的退化将会止步吗？确实，不同生活史细菌基因组的比较分析表明，基因退化最后将会缓慢，至少在专性细胞内寄生菌中是这样。其中有寄主水平对细菌基因功能的筛选[47]。因此，两种细菌从 5000 万年前分化后，已经将基因组锐减到 650kb，但是，剩余部分在结构上近乎完美的稳定，自它们分化后仅丢失了几个基因[47]，对这种极端稳定的解释是：已从基因组中去除了一些重要的重组基因和重复位点，因而减少了进一步由重复序列介导的删除事件[47]；另一种可能性是这些小基因组会侵入到相同生存空间中的拥有更大基因组的细菌所取代。

立克次氏体和线粒体

当选择压存在时，胞内寄生细菌的简并基因组可以在真核细胞内维持几千万年，线粒体是最好的例子。线粒体与真核细胞的共生关系已成为所有高等生物的进化基础，不寻常的是致人死亡的斑疹伤寒病原是与线粒体的祖先最近缘的细菌之一。根据生物能和翻译系统基因重建的种系发生树可以明显地看出，线粒体与包括立克次氏体在内 α 多形菌间的关系[57,58]。

然而，只有很少线粒体蛋白质由线粒体基因组编码，大多数蛋白由核基因组编码。为了研究核基因编码线粒体蛋白的起源与进化，分析了核基因组上 400 多个酵母基因，实验已证明它们都编码线粒体蛋白[59]，从而发现约 50％基因与细菌基因同源，依此推断它们起源于细菌。它们主要与翻译、能量和小分子物质合成有关，种系发生重建确定了这些基因与 α-多形菌之间的密切关系，其余 50％与细菌没有同源性的这些基因，据

推测起源于真核生物，它们主要与膜、转运、调控、信使 RNA 稳定性和剪切有关。

对游离在细胞质中多核糖体的信使 RNA 和线粒体上的信使 RNA 的研究[60]，揭示了非常有趣关于基因起源的现象，推测起源于真核生物的基因，主要在游离细胞质中的多核糖体上翻译，推测起源于细菌的那些基因[59]，都在附着线粒体上的多核糖体上翻译[60]。根据这些相关性可以推测，最初从细菌基因组转移到核基因组上的那些基因产物，是利用细菌的共翻译分泌系统（cotranslational secretion system）转运到线粒体中，一旦这种转运系统建立，由核基因组编码的蛋白也同样能转运回线粒体。

综上所述，编码线粒体中与生物能和翻译过程相关的主要蛋白的基因，是从古老内共生 α-多形菌获得，许多这些基因从线粒体基因组转移到核基因组。然而，大多数现在的线粒体蛋白似乎并不起源于 α-多形菌，而是利用核基因为线粒体服务，因而，现在线粒体的蛋白质组有双重起源：一系列主要蛋白质来自某种古老的 α-多形菌，这种古老 α-多形菌同样也是现在立克次氏体的祖先[61,62]。

结语

对专性细胞内寄生物普氏立克次氏体和康氏立克次氏体的全基因组序列的分析，主要发现这些基因组尽管已经变得很小，但仍然处于衰退状态中。从立克次氏体基因组的序列数据能够从细节上研究基因的失活、退化和切除，基因组的退化可以用多种不同突变机制来解释，每一种都会在向细胞内环境迁移的过程中某一特定时期起主要作用。

早期几个多拷贝基因和重复序列的重组可能造成大量序列的缺失[46,47]，据推测，这个过程伴随基因拷贝数的减少。确实，在其他细菌中为多拷贝的许多基因，在立克次氏体和大多数其他寄生细菌中为单拷贝，例如 rRNA 基因和延长因子基因等[49~52]。接着，发生在短重复序列元件间的重组，引起了相关序列的删除和重复元件的丢失，这样的删除已在不同种的立克次氏体中报道过[41,51,52]。然而，因为越来越多重复序列在此过程中消耗掉，这种事件发生的频率下降，继之而来失活基因的删除，最终可能由复制滑移（replication slippage）诱发在非常短的基因内删除。由于这样删除的核苷酸越来越少，因而退化过程的频率越来越低。

这种最初在立克次氏体中发现的简并进化过程，后来在其他许多专性胞内寄生病原中发现，现在认为麻风分枝杆菌是专性胞内寄生病原中基因大量失活的最好例证[42]。在自由生活细菌的基因组中，DNA 的变迁可能更多，因为它含有更高比例的重复序列和遗传寄生序列[42]，然而，在这些生物中，序列的缺失又可能被相应序列的加入得到补充，同样也可以受益于噬菌体、质粒和其他迁移性遗传元件。

尽管基因退化过程在专性细胞内寄生细菌中最终变慢，许多专性胞内寄生菌会在选择压力下由于进一步退化和小群体的原因而灭绝。另一种由极端寄主特异性带来的威胁是所选择的载体或寄主可能会灭绝，像普氏立克次氏体那样，由于载体虱子的群体数量因不断提高的卫生条件而逐渐下降，细菌病原的生存就没有多少希望。因此，可以想像，普氏立克次氏体在它的基因组退化到最小程度之前，就会在生态环境中消失。

（江　昊　译）

参考文献

1. Moulder JW. Comparative biology of intracellular parasitism. Microbiol Rev 1985; 49:298–337.
2. Andersson SGE, Kurland CG. Reductive evolution of resident genomes. Trends Microbiol 1998; 6:263–278.
3. Winkler HH. Rickettsia species (as organisms). Annu Rev Microbiol 1990; 44:131–153.
4. Olsen GJ, Woese CR, Overbeek R. The winds of (evolutionary) change, breathing new life into microbiology. J Bacteriol 1994; 176:1–6.
5. Andersson SGE, Zomorodipour A, Andersson JO, et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 1998; 396:133–140.
6. Stephens RS, Kalman S, Lammel C, et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 1998; 282:754–759.
7. Ogata H, Audic S, Renesto-Audiffren P, et al. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 2001; 293:2093–2098.
8. Kalman S, Mitchell W, Marathe R, et al. Comparative genomics of *Chlamydia pneumoniae* and *C. trachomatis*. Nature Genetics 1999; 21:395–389.
9. Read TD, Brunham RC, Shen C, et al. Genome sequence of *Chlamydia trachomatis* MoPn *and Chlamydia pneumoniae* AR39. Nucleic Acids Res 2000; 28:1397–1406.
10. Shirai M, Hirakawa H, Kimoto M, et al. Comparison of whole genome sequences of *Chlamydia pneumoniae* J128 from Japan and CWL029 from USA. Nucleic Acids Res 2000; 28:2311–2314.
11. Rachek LI, Tucker AM, Winkler HH, Wood DO. Transformation of *Rickettsia prowazekii* to rifampicin resistance. J Bacteriol 1998; 180:2118–2124.
12. Ricketts HT. JAMA 1909; 52:379–380.
13. Nicolle C, Comte C, Conseil E. C R Acad Sci 1909; 149:486–189.
14. Gross L. How Charles Nicolle of the Pasteur Institute discovered that epidemic typhus is transmitted by lice: reminiscences from my years at the Pasteur Institute in Paris. Proc Natl Acad Sci USA 1996; 93:10,539–10,540.
15. Raoult D, Ndihokubwayo JB, Tissot-Dupont H, et al. Outbreak of epidemic typhus associated with trench fever in Burundi. Lancet 1998; 352:353–358.
16. Hackstadt T. The biology of Rickettsiae. Inf Agents Dis 1996; 5:127–143.
17. Weiss E, Moulder JW. The rickettsias and chlamydias. Order 1. Rickettsiales Giesszckiewicz 1939, 25. In: Krieg NR, Holt JG (eds). Bergeys Manual of Systematic Bacteriology. Baltimore, MD, Williams and Wilkins, 1984, pp. 687–729.
18. Azad AF, Beard CB. Rickettsial pathogens and their arthropod vectors. Emerg Infect Dis 1998; 4:179–186.
19. Raoult D, Roux V. Rickettsioses as paradigms of new or emerging infectious diseases. Clin Microbiol Rev 19:694–719.
20. Moran NA, Munson MA, Baumann P, Ishikawa H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proc R Soc London Ser B 1993; 253:167–171.
21. Roux V, Ridkina E, Eremeeva M, Raoult D. Citrate synthase gene comparison, a new tool for phylogenetic analysis, and its application for the Rickettsiae. Int J Syst Bacteriol 1997; 47:252–261.
22. Andersson SGE, Stothard DR, Fuerst P, Kurland CG. Molecular phylogeny and rearrangement of rRNA genes in *Rickettsia* species. Mol Biol Evol 1999; 16:987–995.
23. Tamura C, Ohashi A, Urakami N, Miyamura S. Classification of *Rickettsia tsutsugamushi* in a new genus, *Orientia* gen nov, as *Orientia tsutsugamushi* comb nov. Int J Syst Bacteriol 1995; 45:589–591.

24. Stothard DR. The Evolutionary History of the Genus *Rickettsia* as Inferred from 16S and 23S Ribosomal RNA Genes and the 17 kilodalton Cell Surface Antigen Gene. Ph.D. thesis. Columbus: Ohio State University, 1995.
25. Walker DH. Rocky Mountain spotted fever: a seasonal alert. Clin Infect Dis 1995; 20:1111–1117.
26. Walker TS. Rickettsial interactions with human endothelial cells in vitro: adherence and entry. Infect Immun 1984; 44:205–210.
27. Gouin E, Gantelet H, Egile C, et al. A comparative study of the actin-based motilities of the pathogenic bacteria *Listeria monocytogenes*, *Shigella flexneri* and *Rickettsia conorii*. J Cell Sci 1999; 112:1697–1708.
28. Heinzen RA, Grieshaber SS, Van Kirk LS, Devin CJ. Dynamics of actin-based movement by *Rickettsia rickettsii* in vero cells. Infect Immun 1999; 67:4201–4207.
29. Walker DH, Firth WT, Edgell CJ. Human endothelial cell culture plaques induced by *Rickettsia rickettsii*. Infect Immun 1982; 37:301–306.
30. Clifton DR, Goss RA, Sahni SK, et al. NF-κB-dependent inhibition of apoptosis is essential for host cell survival during *Rickettsia rickettsii* infection. Proc Natl Acad Sci USA 1998; 95: 4646–4651.
31. Tillier ERM, Collins RA. Genome rearrangement by replication-directed translocation. Nature Genet 2000; 26:195–197.
32. Andersson JO, Andersson SGE. Genome degradation is an ongoping process in *Rickettsia*. Mol Biol Evol 1999; 16:1178–1191.
32a. Tucker A, Winkler HH, Driskell LO, Wood DO. *S*-adenosylmethionine transport in *Rickettsia prowazekii*. J Bacteriol 2003; 185:3031–3035.
33. Katinka MD, Duprat S, Cornillot E, et al. Genome sequence and gene compaction of the eukaryotic parasite *Encephalitozoon cuniculi*. Nature 2001; 414:450–453.
34. Heldt HW. Adenine nucleotide translocation in spinach chloroplasts. FEBS Lett 1969; 5:11–14.
35. Pozueta-Romero J, Frehner M, Viale AM, Akazawa T. Direct transport of ADPglucose by an adenylate translocator is linked to starch biosynthesis in amyloplasts. Proc Natl Acad Sci USA 1991; 88:5769–5773.
36. Schunemann D, Borchert S, Flugge UI, Heldt HW. ADP/ATP translocator from pea root plastids. Comparison with translocators from spinach chloroplasts and pea leaf mitochondria. Plant Physiol (Rock) 1993; 103:131–137.
37. Kampfenkel K, Möhlman T, Batz O, Van Montagu M, Inze D, Neuhaus HE. Molecular characterization of an *Arabidopsis thaliana* cDNA encoding a novel putative adenylate translocator of higher plants. FEBS Lett 1995; 374:351–355.
38. Möhlman T, Tjaden J, Schwöppe C, Winkler HH, Kampfenkel K, Neuhaus H. R. Occurence of two plastidic ADP/ATP transporters in *Arabidopsis thaliana*. Eur J Biochem 1998; 252:353–359.
39. Kuan J, Saier MH. The mitochondrial carrier family of transport proteins: structural, functional. and evolutionary relationships. Crit Rev Biochem Mol Biol 1993; 28:209–233.
40. Cai J, Winkler HH. Transcriptional regulation in the obligate intracytoplasmic bacterium *Rickettsia prowazekii*. J Bacteriol 1996; 178:5543–5545.
41. Tjaden J, Winkler HH, Schwoppe C, Van der Laan MV, Möhlman T, Neuhause HE. Two nucleotide transport proteins in *Chlamydia trachomatis*, one for net nucleoside triphosphate uptake and the other for transport of energy. J Bacteriol 1999; 181:1196–1202.
42. Cole ST, Eiglemeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001; 409:1007–1011.
43. Andersson JO, Andersson SGE. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev 1999; 9:664–671.
44. Andersson JO, Andersson SGE. Pseudogenes, junk DNA and the dynamics of *Rickettsia* genomes.

Mol Biol Evol 2001; 18:829–839.

45. Andersson SGE, Sharp PM. Codon usage and base composition in *Rickettsia prowazekii*. J Mol Evol 1996; 42:525–536.
46. Frank AC, Amiri H, Andersson SGE. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. Genetica 2002; 115:1–12.
47. Tamas I, Klasson L, Canback B, et al. Fifty million years of genomic stasis in endosymbiotic bacteria. Science 2002; 28:2376–2379.
48. Andersson SG, Kurland CG. Genomic evolution drives the evolution of the translation system. Biochem Cell Biol 1995; 73:775–787.
49. Andersson SGE, Zomorodipour A, Winkler HH, Kurland CG. Unusual organization of the rRNA genes in *Rickettsia prowazekii*. J Bacteriol 1995; 177:4171–4175.
50. Syvanen AC, Amiri H, Jamal A, Andersson SGE, Kurland CG. A chimeric disposition of the elongation factor genes in *Rickettsia prowazekii*. J Bacteriol 1996; 178:6192–6199.
51. Andersson SGE, Stothard DR, Fuerst P, Kurland CG. Molecular phylogeny and rearrangement of rRNA genes in *Rickettsia* species. Mol Biol Evol 1999; 16:987–995.
52. Amiri H, Alsmark CM, Andersson SGE. Proliferation and deterioration of *Rickettsia* palindromic elements. Mol Biol Evol 2002; 19:1234–1243.
53. Ogata H, Audic S, Barber V, et al. Selfish DNA in protein coding genes. Science 2000; 290: 347–350.
54. Ogata H, Audic S, Claverie J-M. Response. Science 2001; 291:299–304.
55. Ogata H, Audic S. Abergel C, Fournier P-E, Claverie J-M. Protein coding palindromes are a unique but recurrent feature in *Rickettsia*. Genome Res 2002; 12:808–816.
56. Davids W, Amiri H, Andersson SGE. Small RNAs in *Rickettsia*: are they functional? Trends Genet 2002; 18:331–334.
57. Sicheritz-Ponten T, Kurland CG, Andersson SGE. A phylogenetic analysis of the cytochrome *b* and cytochrome *c* oxidase I genes supports an origin of mitochondria from within the Rickettsiacease. Biochem Biophys Acta 1998; 1365:545–551.
58. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Nature 1999; 283:1476–1481.
59. Karlberg O, Canbäck B, Kurland CG, Andersson SGE. The dual origin of the yeast mitochondrial proteome. Yeast 2000; 17:170–187.
60. Marc P, Mageot A, Devaux F, Blugeon C, Corral-Debrinski M, Jacqu C. Genome-wide analysis of mRNAs targeted to yeast mitochondria. EMBO Rep 2002; 3:159–164.
61. Andersson SGE, Kurland CG. Origins of mitochondria and hydrogenosomes. Curr Opin Microbiol 1999; 5:535–541.
62. Kurland CG, Andersson SGE. Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev 2000; 64:786–820.

18
低G+C含量革兰氏阳性细菌基因组

Steven R. Gill

引言

低G+C含量革兰氏阳性细菌家族包括许多不同的类群，有些是非致病菌，有些是新致病菌，还有一些是人类所知的强致病菌，甚至有些具有极强的杀伤力，成为潜在的生物武器。肠球菌（*Enterococci*）、葡萄球菌（*Staphylococci*）、链球菌（*Streptococci*）和梭状芽孢杆菌（*Clostridia*）等低G+C含量革兰氏阳性细菌，感染人类所引起的发病率和死亡率相当惊人。而且，由于这类病菌对抗生素产生了越来越强的抗性，所以，寻求对这些传染病新治疗方法和控制方法成为人类迫在眉睫的任务。同时，炭疽芽孢杆菌（*Bacillus anthracis*）已被生物恐怖主义所利用，这更增加了人类探究这类病原菌生理和毒理的紧迫性。能引起人类疾病的还有各种支原体，对它们的研究主要将其作为一种模式系统，研究什么是最小基因组或核心基因组，而对它们作为一种慢性病原体的研究相对较少。

对不同低G+C含量革兰氏阳性细菌进行基因组测序的理由不同，就像它们所引起的症状和它们独特的生理特性各不相同一样，对有些细菌进行基因组测序是为了鉴定它们基因组中的致病基因或与它们的致病性有关的基因家族，或是为了探究它们得以在广寄主环境中生存的独特代谢途径；同时，也可以通过基因组测序研究它们基因组间的差异性，因为正是它们各自独特的基因给予了它们特定的致病表型，如抗生素抗性和渐增的毒性。而对另一些菌进行基因组测序，目的是探究它们的代谢途径，从而能将它们更好地用于化学工业和乳品加工业。本章收集了各种低G+C含量革兰氏阳性细菌的基因组测序信息，并对这些信息进行了归纳和总结。

支原体属

支原体（*Mycoplasmas*）属于柔膜体纲（Mollicutes），是目前所知最小的自养微生物[1]，它们的寄主范围很广，包括人类、动物、昆虫和植物。虽然，支原体由低G+C含量革兰氏阳性菌进化而来，但是没有其祖先拥有的刚性G^+细胞壁，而只有富含脂肪胞壁酸和固醇的细胞质膜。在其膜的表面是与其黏附寄主细胞有关的独特终端结构或附属器官[2]。

支原体最显著的特征是其精简的基因组和对密码子的偏爱，其基因组只有约0.5～1.0Mb，识别UGA为色氨酸密码子而不是作为终止密码子[1]。早在20世纪80年代，几位研究人员[3,4]率先提议，可以用支原体定义什么是细胞能够自主复制所必需的最小基因组，这一提议得到许多生物学家的响应，立即开始了对几种支原体基因组测

序[5~8]，并利用转座子突变体库研究支原体的最小基因组[9]。

生殖道支原体和肺炎支原体基因组

生殖道支原体（*Mycoplasma genitalium*）在人的泌尿生殖道中可以引起非淋病性尿道炎，肺炎支原体（*Mycoplasmas pneumoniae*）是人的呼吸道致病菌，可引起儿童和青少年的非典型肺炎。这两种支原体的专性寄居范围不严格，肺炎支原体从泌尿生殖器的临床样品中分离到[10]，而生殖道支原体也从感染肺炎支原体的病人呼吸道中分离到[11]。

生殖道支原体的基因组为580，070 bp环状染色体，平均G+C含量为32%，推测有480个可读框（ORF）[5]。肺炎支原体的基因组大小为816,394 bp，平均G+C含量为40%，推测有677个可读框（ORF）[6,12]。通过对它们基因组的比较发现，肺炎支原体的基因组包含了生殖道支原体的全套基因组。肺炎支原体基因组中另外209个ORF分为两类，一类是与特有功能相对应的编码序列，这是与生殖道支原体存在差异的基础；另一类在生殖道支原体基因组中存在，但在肺炎支原体基因组中有重复或扩增的那些编码序列[12]。通过对这两种支原体的基因组进行基因排序发现，它们的基因组又能下分为6个部分，并且在每部分中基因都按一定的顺序排列，但是，这6个部分的排列顺序在这两种支原体间不一样，好像是这6个部分的位置发生了移动的结果[12]。

在细菌基因组中控制生物复制和转录等基本代谢的基因，一般是比较保守和稳定的，但在支原体基因组中存在细胞生长必需基因的缺失或检测不到的现象[5,6,12]。然而，如果这些关键基因属鉴定不到，或许具有生物学意义，因为，这可能反映这些支原体的某些基因发生了进化分离，以至现今的基因同源性搜索无法检测[13]，例如，认为与DNA聚合酶Ⅲ的装配和形成有关的两个基因 *DNAθ* 和 *DNAδ*，至今仍未找到。另外，在支原体细胞中缺少一般细菌中都有、用于控制基因表达的双组分信号传导系统或其他调节系统[13]。显然，支原体精简的基因组与其代谢途径所需酶的组分和数量的减少相对应，但同时，这要求它们必须在寄主细胞内利用寄主代谢产物生存，如果在体外生存时则需要营养丰富的复杂人工合成培养基[14]。

解脲脲原体基因组

解脲脲原体（又称解脲支原体）（*Ureaplasma urealyticum*）是第三个进行基因组测序的支原体，在分类上属于脲原体属（*Ureaplasma*），寄居于人的泌尿生殖道中，解脲脲原体是妇女妊娠过程中的重要条件致病菌，能引起妇女非淋菌性尿道炎、早产、败血病、脑膜炎以及新生儿肺炎[15]。解脲脲原体的基因组为751,719bp环状染色体，平均G+C含量为25.5%，推测有613个ORF[7]，在这613个基因中，只有324个基因与生殖道支原体和肺炎支原体有高度同源性。而前面提到生殖道支原体的所有基因，在肺炎支原体基因组中均存在对应同源序列，而且，解脲脲原体与生殖道支原体和肺炎支原体的同源序列缺乏共线性，这进一步证实，在基因组进化过程中，解脲脲原体与生殖道支原体和肺炎支原体的基因组发生了进化分离。

在解脲脲原体特有的另外289个基因中，只鉴定了76个基因的功能[7]，它们大多与铁离子获取和通过尿素水解产生ATP有关，而且，解脲脲原体中95%ATP是通过尿

素酶分解尿素产生[16]，这是与其他任何细菌相区别的最明显特征。通过转座突变技术[9]定义的最小基因或核心基因的比较发现，解脲脲原体的基因组，与生殖道支原体基因组中 69 个非必需基因和 255 个必需基因同源[7]。

肺支原体基因组

肺支原体（*Mycoplasma pulmonis*）与海洋生物的呼吸道支原体病有关，也是研究支原体引起呼吸道传染病的最好模型[17]。肺支原体的基因组为963,879bp 环状染色体，平均 G+C 含量为 26.6%，推测有约 782 个 ORF，是已完成基因组测序的所有支原体中最大的[8]。与已完成基因组测序的其他柔膜体纲支原体相比，肺支原体大部分特有基因是编码运输蛋白和膜蛋白[8]。肺支原体基因组最显著的特征是有多种产生抗原多样性的机制，其中包括：①在可变表面抗原基因座位（*vsa* locus）上膜表面脂蛋白基因的多样性表达，而产生的相变化[18]；②膜表面蛋白基因上游高度重复序列 DNA 链滑动，而产生的碱基错配[8]。如果实验能证实抗原多样性产生的这两种机制，将会第一次表明在一个支原体细胞中有两种不同机制，控制膜表面抗原基因的相变表达。

最小基因组

定义什么是最小基因组需要借助先进的计算机技术和现代实验方法。将流感嗜血菌（*Haemophilus influenzae*）与生殖道支原体的基因组通过计算机比较分析发现，两种细菌基因组共有 256 个基因是它们细胞生存和自我复制所必需的，称为核心基因组[19]。最近，科学家运用生殖道支原体和肺炎支原体的转座子突变体库，来定义支原体在体外生存时所必需的最小基因组[9]，他们的方法是将转座子 Tn4001 插入到生殖道支原体和肺炎支原体的染色体中，并在体外培养，如果在非必需基因内或在不同基因之间的非编码区域插入转座子，并不能导致细菌细胞死亡。通过这种方法，他们确定支原体在体外复制所必需的核心基因约有 265～350 个，这些核心基因编码 DNA 复制和基因转录所必需的蛋白质、细胞从环境中吸取营养所必需的转运蛋白质、细胞产生 ATP 和还原力所必需的酶类以及细胞保持内环境稳定所必需的组分。但是，在体内寄生时，菌体需要吸附在寄主组织上，而且缺乏核酸和蛋白质等代谢物，其所必需的基因数似乎要多一些。

正在进行的支原体基因组计划

还有一些支原体正在进行基因组测序，它们是丝状支原体丝状亚种 SC（*M. mycoides* subsp. *mycoides SC*）、猪肺炎支原体（*M. hyopneumoniae*）、山羊支原体（*M. capricolum*）、关节炎支原体（*M. athritidis*）、鸡毒支原体（*M. gallisepticum*）、穿透支原体（*M. penetrans*）、短吻鳄支原体（*M. alligatoris*）、柑橘僵化病螺原体（*Spiroplasma citri*）和玉米矮缩病螺原体（*Spiroplasma kunkelii*）[8]，如果这些支原体的基因组测序能顺利完成，它们将会成为不同基因组间相互比较和微生物进化研究很好的实验模型，并能在此基础上对最小基因组作进一步描述。

李斯特菌属

产单核细胞李斯特菌（又称单核细胞增生李斯特菌）基因组和无害李斯特菌基因组

李斯特菌属（*Listeria*）是一类革兰氏阳性兼性厌氧杆状细菌，广泛存在于土壤、水、食物及人和动物的粪便中[20]。致病性李斯特菌有两种：一种是产单核细胞李斯特菌（*L. monocytogenes*），能感染人和其他脊椎动物；另一种是伊氏李斯特菌（*L. ivanovii*），能引起牛羊的早产、死胎或导致新生畜感病[20]，但极少感染人类[21]。产单核细胞李斯特菌是引起李斯特菌病的主要致病菌，能通过食物传播，常导致严重局部性或普遍性感染，如脑膜炎、流产、败血病、胃肠炎和妇女围产期传染病[20]。在对产单核细胞李斯特菌进行基因组测序前，曾推测其基因组中一段10kb DNA序列可能编码致病蛋白因子，因为，在同属非致病菌无害李斯特菌的基因组中是没有这一段序列的[22]，所以，无害李斯特菌常常作为表达产单核细胞李斯特菌基因的寄主菌[23]。产单核细胞李斯特菌的致病因子包括：(1) 位于细胞表面的内化素蛋白（*internalin*）In1A，它是菌体穿透寄主肠壁上皮细胞所必需的；(2) 侵入蛋白InlB；(3) 能使菌体从吞噬细胞液泡中逃离的LLO和PICA；(4) 为细胞内肌动蛋白运动和细胞间扩散提供工具的ActA和PicB[20,24]。

产单核细胞李斯特菌血清型1/2a菌株EGD-e的基因组大小为2 944 528bp，平均G+C含量为39%，推测可编码2853个基因；无害李斯特菌血清型6a菌株CLIP11262的基因组大小为3 011 209bp，平均G+C含量为37%，推测可编码2973个基因。另外，这两种细菌的基因组都含有原噬菌体序列，无害李斯特菌还有一个81 905bp的质粒。而且，这两种李斯特菌的基因组间具有高度保守性和共线性，它们与枯草芽孢杆菌（*Bacillus subtilis*）和金黄色葡萄球菌（*Staphylococcus aureus*）的基因组间也有共线性[25,26,27]。与无害李斯特菌的基因组相比，产单核细胞李斯特菌基因组还含特有的270个基因，它们分散在基因组许多长达1～25kb不同区段中。这些基因的分散排列类似埃希氏大肠杆菌（*Escherichia coli*）O157：H7和K12基因组[28]，但不同于链球菌（*Streptococcus*）基因组，链球菌基因组因原噬菌体序列而产生了多样性[29]。

通过对产单核细胞李斯特菌血清型1/2a菌株EGD-e的基因组和无害李斯特菌血清型6a菌株CLIP 11262的基因组进行比较分析，鉴定到产单核细胞李斯特菌特有编码致病蛋白因子的基因，其中包括各种分泌蛋白基因、脂肪酶基因和几丁质酶基因[22]。产单核细胞李斯特菌的基因组能编码41种以上特殊的细胞表面蛋白，这些细胞表面蛋白含有分选酶所必需的LPXTG细胞壁连接域[30]。而在其他已完成测序的革兰氏阳性细菌基因组中，能编码这些表面蛋白质的种类要少的多，如化脓链球菌（*S. pyogenes*）基因组只能编码13种，金黄色葡萄球菌基因组只能编码18种这样的细胞表面蛋白[27,31,32]。在这些细胞表面的内化素蛋白家族中，从产单核细胞李斯特菌和无害李斯特菌中，分别找到19种和8种含LPXTG细胞壁连接域的蛋白[22,30]。

在产单核细胞李斯特菌基因组和无害李斯特菌基因组中，含有几乎相同数量的转录调节子，前者是209个，后者是203个，而且，这两种李斯特菌基因组都编码相同数量

的双组分调节系统[22]。这反映了它们需要对各自不同生存环境做出不同适应性反应，不过无害李斯特菌基因组并不编码产单核细胞李斯特菌所特有的多种致病因子[20,33]，类似这样调节系统的平衡，也同样存在于强致病性金黄色葡萄球菌和弱致病性表皮葡萄球菌（*S. epidermidis*）之间[27]。与无害李斯特菌相比，产单核细胞李斯特菌独有的调节子是基本调节子 A（PrfA），是它激活了产单核细胞李斯特菌多种致病基因的表达[20,33]。

目前，另外两种正在进行基因组测序的李斯特菌是伊氏李斯特菌（*L. ivanovii*）和产单核细胞李斯特菌血清型 4b 菌株，伊氏李斯特菌是反刍类动物的致病菌，与牛羊等牲畜的流产、死胎和新生畜的败血症有关[20]，产单核细胞李斯特菌血清型 4b 菌株曾引起多起重大食物传染性疾病的爆发流行[34,35]。

肠球菌属

肠球菌属（*Enterococci*）是一类革兰氏阳性、兼性厌氧的球状细菌，它们大多共生于健康的人和动物的肠道内，也广泛存在于土壤、下水道、水和食物中。在美国，它是抗生素抗性和医院获得性感染的主要原因之一[36]，有两种肠球菌能引起人传染病，它们是粪肠球菌（又称粪链球菌）（*Enterococcus faecalis*）和屎肠球菌（又称屎链球菌）（*Enterococcus faecium*），但 80％人类肠球菌传染病是由粪肠球菌引起的[37]，这或许是因为在人肠道中粪肠球菌比屎肠球菌数量多的缘故，也可能是粪肠球菌比屎肠球菌有更强的致病性[37]。万古霉素一直是人类控制抗生素抗性剧增 G^+ 致病菌流行最有效的抗生素类药物，但近来发现，许多 G^+ 致病菌对盐酸万古霉素也产生了耐药性，这给人类治疗肠球菌引起人类传染病增加了更大的困难[38,39]。粪肠球菌菌株 V583 基因组测序已经完成，屎肠球菌基因组测序已达到了 8 倍的基因组覆盖率(http://www.jgi.doe.gov/index.html)。

粪肠球菌基因组

粪肠球菌（*Enterococcus faecalis*）菌株 V583 是临床分离的第一个对万古霉素产生抗性的肠球菌[38]，基因组大小为3 218 030bp，G+C 含量为 38％，推测可编码3,490个基因[40]。通过与已完成测序的其他细菌基因组进行比较发现，粪肠球菌基因组中有至少 85％基因与低 G+C 含量的其他革兰氏阳性细菌的基因组同源。另外，粪肠球菌含有 3 个大小在17 963～66 320bp 之间的环状质粒已完成测序，其中 2 个质粒在结构上与感应信息素质粒 pAD1 和 pCF10 有相似性，第 3 个质粒则是广寄主范围质粒家族 pAMβ1 的成员[41～43]。

粪肠球菌基因组中有 7 个区域（占整个基因组约 10.2％）是原噬菌体序列，与其他低 G+C 含量革兰氏阳性细菌基因组中的噬菌体序列有很紧密的亲缘关系，这与 A 族链球菌基因组中的原噬菌体序列很相似[29]，这表明，肠球菌和链球菌的进化关系很近。粪肠球菌菌株 V583 基因组中有 150kb 的一大段序列（称为致病基因岛）[44]编码致病因子，其中包括胆汁酸水解酶和众所周知的细胞表面定殖蛋白 Esp。在粪肠球菌菌株 MMH594 和 V586 中也存在几乎相同的致病基因岛，但是，它们的致病基因岛中含有其

他插入序列（称为IS元件），而且在菌株MMH594的致病基因岛中，还插入2.8kb的溶细胞素操纵子[44]，在粪肠球菌中已鉴定了一些功能类似黏附素或聚合因子的细胞表面蛋白。值得注意的是，在许多推定编码细胞表面蛋白的基因内，含有许多同聚或交互重复的核苷酸模体（motif），这可能导致细胞表面蛋白相变表达理论中的链滑动错配机制，类似的链滑动机制也存在其他病原微生物中，如肺炎链球菌（*Streptococcus pneumoniae*）就是利用这一机制进行细胞表面抗原基因的相变表达[45]。

粪肠球菌菌株V583对万古霉素的抗性与两个截然不同的可移动元件有关，一个元件是包含万古霉素抗性基因 *vanA* 的一段DNA序列，来自粪肠球菌接合转座子Tn1546[46]。万古霉素抗性基因 *vanA* 的表达，导致粪肠球菌菌株V583对万古霉素的高水平抗性。抗万古霉素金黄色葡萄球菌（VRSA）首次从两位患者中发现，可能是因为接合转移使 *vanA* 基因从抗万古霉素粪肠球菌（VRE）的基因组中，通过转座作用插入到金黄色葡萄球菌基因组中。另一个元件是一段与肺炎链球菌菌株VncRS基因组中编码双组分信号传导系统的基因高度同源的序列[47]，这段序列的表达增强了粪肠球菌菌株V583对万古霉素的耐药性[48]。

葡萄球菌属

葡萄球菌属（*Staphylococci*）是一类革兰氏阳性、兼性厌氧的球状细菌，能引起人和动物的多种疾病，其中有的是一般性常见病，如轻微的皮肤传染病和食物中毒，也有的是严重的甚至致命的疾病，如心内膜炎、中毒性休克综合征和脑膜炎。尽管人类花费了大量的精力来控制葡萄球菌的蔓延，但它在全球仍然不仅是医院还是社区获得性感染的罪魁祸首[49~51]。

金黄色葡萄球菌和表皮葡萄球菌是葡萄球菌属中典型的条件致病菌[50~52]，它们主要共生于鼻粘膜表皮细胞和皮肤表皮细胞，一般对寄主细胞不造成伤害。然而，当这些表皮细胞受外界创伤时，如外伤、接种、输液或手术等，它们就会进入寄主细胞成为致病菌。

表皮葡萄球菌主要通过外来物（如医疗器械）携带传染，金黄色葡萄球菌是一类更具传染性的致病菌，常引起急性和化脓性感染[50~52]。抗甲氧苯青霉素金黄色葡萄球菌（MRSA）和抗甲氧苯青霉素表皮葡萄球菌（MRSE）对甲氧苯青霉素的耐药性，以及中度抗万古霉素金黄色葡萄球菌（VISA）对万古霉素的中度抗性，使原来有效控制和治疗葡萄球菌所引发疾病的方法面临新挑战[48,53]。除了面对医院获得性抗甲氧苯青霉素金黄色葡萄球菌（H-MRSA）和中度抗万古霉素金黄色葡萄球菌（H-VISA）的威胁外，传染病临床医生必须面对来自新出现的社区获得性抗甲氧苯青霉素金黄色葡萄球菌（C-MRSA）和抗甲氧苯青霉素敏感金黄色葡萄球菌（C-MSSA）的更大威胁，因为它们要比典型的医院获得性抗甲氧苯青霉素金黄色葡萄球菌（H-MRSA）的毒性更高[54~56]。

虽然，对金黄色葡萄球菌和表皮葡萄球菌的生理和毒性机制已了解得很多，但仍要对它们的基因组进行测序，以便在比较分析它们基因组间差异性的基础上，进一步探明那些至今未知的毒性机制和它们对甲氧苯青霉素和万古霉素产生抗药性的机制。在全球

众多研究中心的通力合作下，葡萄球菌 7 个菌株的基因组测序已经完成（www.tigr.org），它们是：① 医院获得性抗甲氧苯青霉素金黄色葡萄球菌（H-MRSA）菌株 COL[27]和 N-315[32]；②来自英国医院的流行性抗甲氧苯青霉素金黄色葡萄球菌菌株 E-MRSA-16（菌株 252）[55]；③医院获得性抗甲氧苯青霉素兼中度抗万古霉素金黄色葡萄球菌（H-MRSA/H-VISA）菌株 Mu50[32]；④实验室菌株 NCTC8325[57]；⑤社区获得性抗甲氧苯青霉素金黄色葡萄球菌（C-MRSA）菌株 MW-2[58]；⑥社区获得性甲氧苯青霉素敏感金黄色葡萄球菌（C-MSSA）高毒菌株 476[55]；⑦抗甲氧苯青霉素表皮葡萄球菌（MRSE）分离株 RP62A[27]。

金黄色葡萄球菌基因组

金黄色葡萄球菌（*Strephylococcus aureus*）不同菌株的基因组大小为 2.8～2.9Mb，但均为环状染色体，平均 G+C 含量为 30%，推测可编码近 2600 个基因[27,32,58]。金黄色葡萄球菌基因组中，80%基因序列与其他低 G+C 含量革兰氏阳性细菌的基因组同源。金黄色葡萄球菌不同菌株的共同特点是基因组中有多种 IS 元件、原噬菌体序列和致病基因岛等可移动元件[59]，这些可移动元件的插入，使金黄色葡萄球菌不同菌株获得不同致病基因和抗生素抗性基因，这些基因又能在不同金黄色葡萄球菌菌株间转移，导致不同菌株基因组间的差异性和高毒菌株的产生[59,60]。典型例子是金黄色葡萄球菌的致病基因岛，包括 SaPl1[60]和 SaPln2，SaPl1 序列包含中等毒性休克综合征基因 *tsst*，SaPln2 序列包含葡萄球菌类外毒素（SET）基因簇[32]。社区获得性抗甲氧苯青霉素金黄色葡萄球菌（C-MRSA）高毒菌株 MW-2 和甲氧苯青霉素敏感金黄色葡萄球菌（C-MSSA）高毒菌株 476 的基因组中，包含 4 个特殊的致病基因岛，这在其他金黄色葡萄球菌 5 个已测序菌株中并未发现过，其中的一个致病基因岛含有特殊肠毒素等位基因 *sel2* 和 *sec4*，这可能是菌株 MW-2 和菌株 476 毒性高的原因[58]。

葡萄球菌的抗生素抗性基因，通常位于质粒或其他可移动遗传元件上[61]，如甲氧苯青霉素抗性基因 *mec* 位于葡萄球菌称为盒式染色体（SSC*mec*）的可移动元件上，该元件整合在基因组复制原点附近[62,63]，它的四种等位形式：Ⅰ型、Ⅱ型、Ⅲ型和Ⅳ型，在葡萄球菌不同菌株中都发现过[62,63]，其大小在 24～100kb 之间不等，结构上也有明显差异。中度抗万古霉素金黄色葡萄球菌（VISA）对万古霉素的最低抑制浓度（MIC）为 8μg/ml，其抗性机制比较复杂，涉及多种生化代谢途径。

虽然，中度抗万古霉素金黄色葡萄球菌（VISA）的双组分调节因子 *vraSR*，在抗万古霉素机制中发挥重要作用[26]，但是，与其细胞壁合成和装配有关的几种代谢途径，也可能参与到对万古霉素的抗性机制中[64]。美国的一家医院 2002 年首次从感染金黄色葡萄球菌的两个病人中，分离到高抗万古霉素的金黄色葡萄球菌（VRSA）两个菌株[48]，它们抗万古霉素的最低浓度（MIC）达 32μg/ml 以上。高抗盐酸万古霉素金黄色葡萄球菌（VRSA）的抗性机制，不同于中度抗万古霉素金黄色葡萄球菌（VISA），因为在首次分离的高抗万古霉素金黄色葡萄球菌（VRSA）两株菌株的基因组中，鉴定到 *vanA* 基因，这是高抗万古霉素肠球菌（VRE）的抗性基因[39]，该基因可能是通过接合作用从高抗万古霉素肠球菌（VRE）基因组，转移到高抗万古霉素的金黄色葡萄

球菌（VRSA）基因组中。

金黄色葡萄球菌基因组可编码多种致病蛋白因子，其中包括各种蛋白酶、降解酶、肠毒素、外毒素、溶血素和一些起黏附寄主作用的表面蛋白[49]，在金黄色葡萄球菌的基因组中至少鉴定了70多种新致病蛋白基因，其中许多致病蛋白基因位于致病基因岛中。这些致病基因编码的一种致病蛋白因子大小约为1kDa黏附蛋白Ebh，可能是一种内皮细胞黏附素[65]或纤维素结合蛋白[66]，在菌体感染人类并引发心内膜炎疾病的过程中发挥一定作用[32]，在表皮葡萄球菌基因组中也证实存在类似黏附蛋白[27]。

金黄色葡萄球菌致病蛋白因子的表达受三方面调控：附属基因调节子（*agr*）[67]、与附属基因调节子（*agr*）相关的双组分信号系统以及葡萄球菌附属调节子（*sar*）家族[68,69]。在金黄色葡萄球菌基因组中总共有16个类似附属基因调节子（*agr*）的双组分信号传导基因对和10个葡萄球菌附属调节子（*sar*）家族成员[27,32]。总之，金黄色葡萄球菌基因组中的原噬菌体序列、致病基因岛和其他可移动元件都可能增加菌体细胞的适应性，使菌体能适应多种寄主环境并抵抗杀菌剂的攻击。

表皮葡萄球菌基因组

表皮葡萄球菌（*Staphylococcus epidermidis*）菌株RP62A基因组为2 619 000bp的环状染色体，平均G+C含量为30%，推测可编码2586个基因[27]，与金黄色葡萄球菌一样，表皮葡萄球菌基因组中有大约80%基因与其他低G+C含量革兰氏阳性细菌的基因组高度同源。在该菌株的基因组中，也鉴定到一段原噬菌体序列，与枯草芽孢杆菌基因组中的原噬菌体序列SPP1类似[70]，在该基因组中还含大小为28kb能编码β-内酰胺酶的质粒，从而使其对氨苄青霉素产生抗性，与金黄色葡萄球菌不同的是，在表皮葡萄球菌基因组中没有致病基因岛。

在表皮葡萄球菌和金黄色葡萄球菌基因组中，都含有一段保守的共线性序列，该序列位于大小为2.2Mb的高度保守区段上，在基因组进化过程中，由于一些原噬菌体序列或其他可移动元件的插入，使这个2.2Mb高度保守区段在这两种葡萄球菌基因组间发生了进化分离[27]，在该区段之外，两种葡萄球菌基因组都含有各自的基因组复制起点，而且发生了更明显的进化分离。

在表皮葡萄球菌编码的2586个基因中，有2067个基因在金黄色葡萄球菌基因组内存在对应基因，这些基因在这两种葡萄球菌中执行基本或核心的功能。有趣的是，表皮葡萄球菌基因组中的可移动元件SSC*mec*，与金黄色葡萄球菌菌株N315的Ⅱ型可移动元件SSC*mec*极为相似，这证明了我们收集数据所表明的结果，即可移动元件SSC*mec*能在多种葡萄球菌基因组间移动[62,63]。

与李斯特菌属中不同致病性细菌基因组间的差异类似，通过表皮葡萄球菌和金黄色葡萄球菌基因组的比较分析表明，这两种葡萄球菌在毒力方面的差异，主要由于表皮葡萄球菌基因组中，缺乏能编码内毒素和外毒素等致病因子的基因。尽管如此，表皮葡萄球菌基因组中，仍含有与葡萄球菌附属调节子（*sar*）、附属基因调节子（*agr*）和类似附属基因调节子（*agr*）的双组分信号传导系统基因等近乎相同的基因序列，这些调节子序列可能与金黄色葡萄球菌和表皮葡萄球菌中核心功能基因的调节有关，同时，也起着适应不同环境刺激的作用。

除了金黄色葡萄球菌和表皮葡萄球菌外，至少有另外两种葡萄球菌正在进行基因组测序，一种是无致病性肉葡萄球菌（*Staphylococcus carnosis*），另一种是能感染牛的金黄色葡萄球菌分离株[71]。通过将感染人和牛的金黄色葡萄球菌基因组进行比较分析，可以鉴定到与致病寄主有关的特征基因和种的特征性致病基因。

芽孢杆菌属

芽孢杆菌属（*Bacillus*）是一类革兰氏阳性，好氧或兼性厌氧，能形成芽孢的杆状细菌。属中的各种细菌形态多样，致病性也不同[72]，多数细菌是土壤腐生菌，有些是人和动物的致病菌，还有些是有用的工业微生物。研究最多的是枯草芽孢杆菌，它是典型的革兰氏阳性细菌，也是典型的由营养细胞产生内生孢子的细菌[73]。枯草芽孢杆菌常用于研究细菌分化和染色体向子细胞分离[74]，此外，由于它能高效分泌大量细胞外蛋白，使其成为 G^+ 细菌研究胞外蛋白分泌和产生外源蛋白的模型[75]。炭疽芽孢杆菌是人和动物的致病菌，被公认为是 21 世纪初重要的细菌病原体，已经成为基因组法医学的应用模型[76]。首次进行基因组测序的三种芽孢杆菌是枯草芽孢杆菌（*B. subtilis*），耐盐芽孢杆菌（*B. halodurans*）和炭疽芽孢杆菌（*B. antherous*）。

枯草芽孢杆菌基因组

对枯草芽孢杆菌基因组测序时遇到了前所未有的技术性困难，其中，最大困难是在大肠杆菌中构建不了枯草芽孢杆菌整个基因组的质粒测序文库，为了解决这些技术难题，科学家们创造性的运用低拷贝克隆载体[77]和聚合酶链反应等技术[78]，达到了理想效果，更重要的是这些技术在以后的基因组测序中得到非常广泛的应用。

枯草芽孢杆菌菌株 168 的基因组为4 214 810bp 环状染色体，平均 G + C 含量为 43.5%，推测可编码4100个基因[25]，在与大肠杆菌基因组比较中发现，它们的基因组中大部分基因具有相似功能，而且有推定的约 100 个操纵子在它们的基因组间高度保守[25]。在与生殖道支原体基因组的比较中发现，在它们的基因组中有 300 个基因高度同源，其中一些基因还保持了类似支原体基因组的最小结构。

枯草芽孢杆菌基因组虽然能编码基因调节、双组分信号传导系统、群体感应（quorum sensing）、蛋白分泌和芽孢形成等所必需的许多功能性组分，但是，却缺乏大肠杆菌基因组所编码的一些基本组分。有趣的是，在大肠杆菌染色体分离中起关键作用的 MukB 蛋白[79]，在枯草芽孢杆菌中却未能找到。然而，枯草芽孢杆菌的 Smc 蛋白与 MukB 蛋白略微有些相似[79,80]，或许能执行类似的功能，这也可能表明，革兰氏阳性细菌和革兰氏阴性细菌染色体的分离机制有微小差别。枯草芽孢杆菌基因组特有的基因，包括与植物（如瘿碱）来源大分子降解有关的酶基因，这些基因或许是枯草芽孢杆菌与其生存环境（土壤和植物）有关系的反映。枯草芽孢杆菌基因组中含有多种原噬菌体序列，其中包括 SPβ、PBSX 和 skin 原噬菌体或隐秘质粒，所有这些可移动遗传元件，在基因转移和基因组多样性发生的过程中发挥重要作用。

耐盐芽孢杆菌基因组

耐盐芽孢杆菌是一类嗜碱细菌，最适 pH 值大于 9.5[81]，对它的研究兴趣主要有两大领域：其一，碱性环境适应机制；其二，商业产品的持续开发，如蛋白酶和纤维素酶等。耐盐芽孢杆菌菌株 C-125 的基因组测序计划已经开始执行，希望通过这一研究能发现嗜碱枯草杆菌的特异功能，并能通过与枯草芽孢杆菌基因组的比较分析，确定不同芽孢杆菌的共同功能。

耐盐芽孢杆菌的基因组为4 202 353bp 环状染色体，平均 G+C 含量为 43.7%，推测可编码 4066 个基因。将耐盐芽孢杆菌基因组和枯草芽孢杆菌基因组进行比较发现，大约有 1500 个管家基因位于两种细菌基因组共有的 DNA 区段上[82]，与枯草芽孢杆菌菌株 168 基因组相比，耐盐芽孢杆菌基因组中没有完整的原噬菌体序列，而枯草芽孢杆菌基因组中至少有 3 个完整的原噬菌体序列，而耐盐芽孢杆菌基因组中有至少 15 个独特的 IS 元件，这些元件主要插入在基因组的非编码区域[83]，在耐盐芽孢杆菌中还鉴定到一些与嗜碱特性有关的蛋白因子，其中包括许多独特的 σ 因子，它们可能调节和控制着菌体在碱性环境中基因的转录。

炭疽芽孢杆菌基因组

炭疽芽孢杆菌是芽孢杆菌属中最具毒力的一种，是典型的土壤栖居菌，能在土壤中形成芽孢并传染给人和动物，经常引起致命疾病——炭疽病。在实验很容易培养并提纯炭疽芽孢杆菌的芽孢，这使其具有被生物恐怖主义用作大规模杀伤性生化武器的潜在能力。最早期对炭疽芽孢杆菌的研究，如巴斯德将其用作疫苗接种绵羊以预防炭疽热病[84]，极大地促进了细菌学和免疫学发展。

炭疽芽孢杆菌是蜡状芽孢杆菌群中的一种，该群还包括苏云金芽孢杆菌（*B. thuringiensis*）和蜡状芽孢杆菌（*B. cereus*）。从系统发生角度看，蜡状芽孢杆菌群与另两种已测序芽孢杆菌—耐盐芽孢杆菌[82]和枯草芽孢杆菌[25]有所不同。炭疽芽孢杆菌已知的毒性基因，包括炭疽毒素基因和聚-D-谷氨酸荚膜基因，它们位于两个已完整测序的质粒 pXO1（181Kb）和 pXO2（93.5Kb）上[85,86]。另外一些蛋白因子基因则至今仍未找到，所以，它们的功能还不得而知。对炭疽芽孢杆菌基因组进行测序的目的，是为了鉴定其基因组中是否存在新潜在毒性基因，并在此基础上开发出能预防炭疽病的潜在疫苗。致病性佛罗里达炭疽热分离株，是第二个进行基因组测序的炭疽芽孢杆菌，测序的目的是为了鉴定其独特基因序列的多态性，并能使其发展成追踪炭疽热爆发流行的新式法医工具[76]。

第一个测序的炭疽芽孢杆菌是来自英格兰威尔特郡波尔特高地（Porton Down），缺失了质粒 pXO1 和 pXO2 的毒性菌株 Ames，该菌株的基因组大小为5 227 297bp 环状染色体，平均 G+C 含量为 35.4%，推测可编码 5753 个基因[87]。炭疽芽孢杆菌的基因组中没有 IS 元件，但是含有 4 个原噬菌体序列，这些序列占全基因组的 2.8%（图 1），这些原噬体序列不编码毒素，但却能编码分泌型蛋白和细胞表面蛋白，这些蛋白可能在菌体与寄主细胞外环境的交互作用中起一定作用。

令人惊奇的是，炭疽芽孢杆菌基因组中大部分毒性基因，在蜡状芽孢杆菌菌株

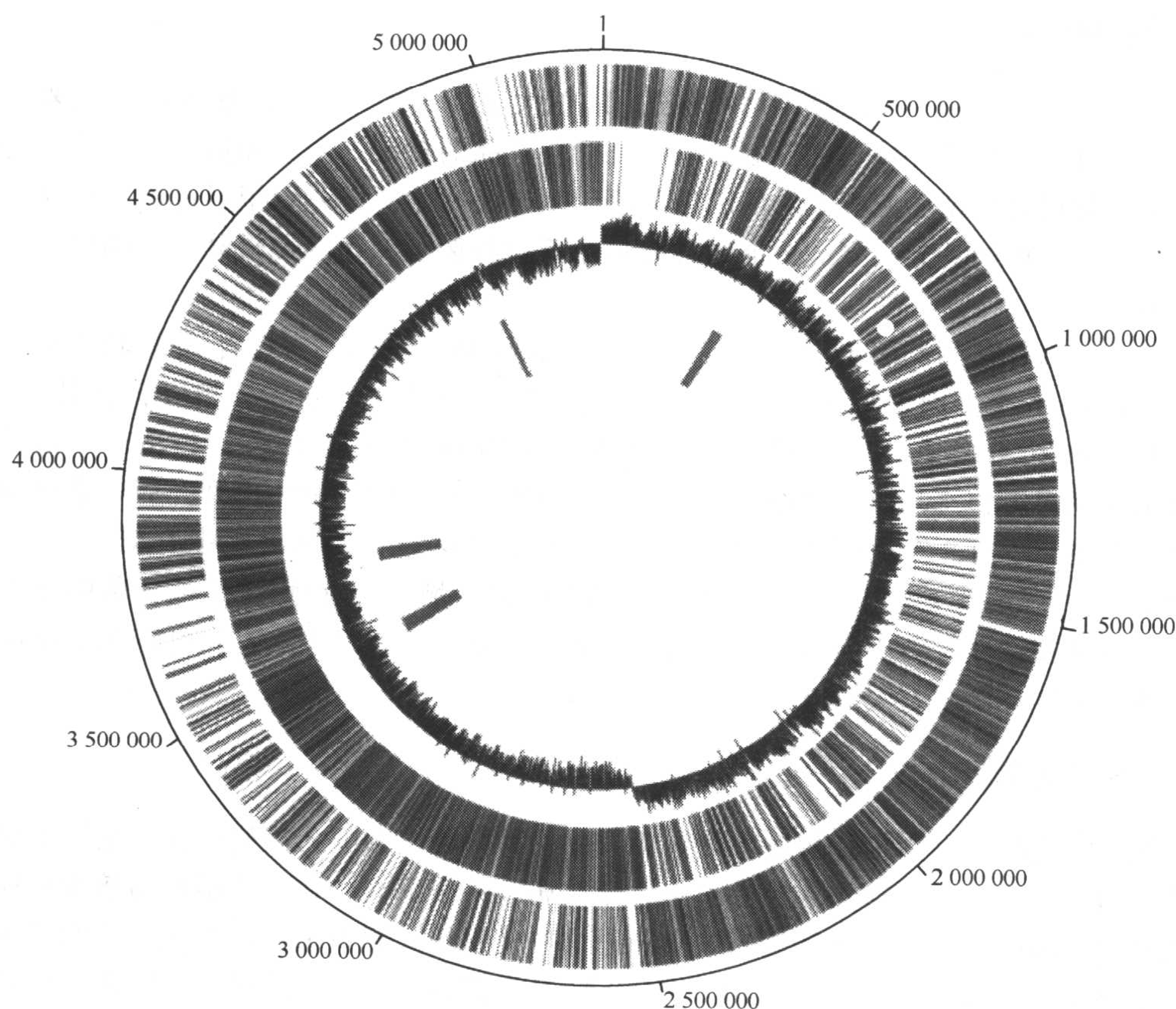


图 1　炭疽芽孢杆菌菌株 Ames 基因组（图中最外圈和第二圈示出了基因的位置和方向，第三圈示出 GC 偏衡（GC skew）情况，最内圈示出 4 个原噬菌体序列在基因组中的位置）。在该菌株基因组中 74％编码序列的转录方向与其复制方向相同，表现出很强的方向偏嗜性。基因的这种方向偏嗜性是所有低 G＋C 含量 G^+ 细菌的共同特征。该菌株的 4 个原噬菌体序列，在其他炭疽芽孢杆菌基因组中也存在，它们能编码多种表面蛋白。（本图来源于 Read et al. 的未发表资料）

10987 的基因组有同源物（蜡状芽孢杆菌基因组计划；www.tigr.org），而这些毒性基因不位于致病基因岛中。虽然，炭疽芽孢杆菌基因组预测可编码多种胞外蛋白，但实验数据[88]表明，这些蛋白质没有分泌到细胞外，可能是在基因组的正调节子 *plcR* 基因上发生了移码突变，该调节子在蜡状芽孢杆菌和苏云金芽孢杆菌中正调控多种胞外蛋白的表达[89]。对炭疽芽孢杆菌细胞表面蛋白基因的研究表明，至少有 29 种表面蛋白可作为潜在疫苗进行开发和利用。对该菌基因组进行分析的最大成果是发现了炭疽芽孢杆菌基因组并不编码自己独特的毒性因子，编码的所有潜在毒性蛋白在蜡状芽孢杆菌菌株 10987 中也鉴定到，这进一步证实了这种观点，即炭疽芽孢杆菌、蜡状芽孢杆菌和苏云金芽孢杆菌都属同一个种，种的特异毒性基因在基因组进化过程中，通过染色体和质粒 DNA 交换而在它们的自然生境中相互转移[87,90]。

目前，基因组研究所正在开展一项更大的炭疽芽孢杆菌基因组测序（www.tigr.org），主要任务是对炭疽芽孢杆菌多个不同分离株进行全基因组测序，并将它们的单核

苷酸多态性（single-nucleotide polymorphisms）进行比较，主要目的是确定各分离株的归属关系，并发展和改进研究炭疽热爆发流行的法医工具。在 Read 等[76]的第一份研究报告中，通过比较佛罗里达炭疽热分离株和菌株 Ames 的基因组，鉴定出 60 个新标记物，其中包括单核苷酸多态性、插入或缺失序列、串联重复等，他们的研究促进了诸如 MUMmer[91]等比较分析工具的发展和改进，这些工具对分析其他基因组也十分有用。

梭状芽孢杆菌

梭状芽孢杆菌（*Clostridia*）是革兰氏阳性，能形成芽孢的杆状厌氧细菌，包括两大类群：一类是能产生毒素的致病菌，如艰难梭菌（*C. difficile*）、肉毒梭菌（*C. botulinum*）、破伤风梭菌（*C. tetani*）以及产气荚膜梭菌（*C. perfiringens*）。另一类是常规种类，如丙酮丁醇梭菌（*C. acetobutylicum*），这类细菌已在化学工业中开发应用，通过发酵生产丙酮、丁醇和其他有机化合物[92,93]。已完成两种梭状芽孢杆菌基因组测序，它们是丙酮丁醇梭菌和产气荚膜梭菌，丙酮丁醇梭菌在 20 世纪早期就通过丙酮—丁醇—乙醇发酵途径生产丙酮[94,95]，近来又重新用于石油化工通过低能耗和低成本生产化学溶剂[95,96]，对丙酮丁醇梭菌菌株 ATCC824 进行基因组测序的兴趣是构建该菌基因工程菌，并可用于开发能生产更多化学溶剂的工程菌[97,98]。产气荚膜梭菌广泛存在于土壤和下水道中，也是人和动物肠道的正常菌群[99]，它的孢子或营养细胞能通过伤口传染寄主细胞，造成寄主组织大面积坏死，通常引起寄主死亡[100]。

丙酮丁醇梭菌基因组

丙酮丁醇梭菌典型菌株 ATCC824 的基因组大小为3,940,880bp，平均 G+C 含量为 30%，推测有 3740 个可读框（ORF）[101]。除染色体 DNA 外，菌株 ATCC824 还有 2 个隐秘噬菌体 DNA 和一个 200kb 质粒 pSOL1，该质粒以前证明能编码与溶剂产生有关的许多基因[102]，进一步分析表明，它有 178 个可读框（ORF），与菌体细胞产生丁醇、乙醇、丙酮等有机溶剂的机制有关。通过与其他细菌基因组比较表明，丙酮丁醇梭菌基因组与枯草芽孢杆菌之间有亲密进化关系，它们之间的主要差别是，枯草芽孢杆菌基因组中含有数目不等与芽孢形成有关的大片段基因，其中包括许多与芽孢形成和芽孢结构有关的调节基因[102]。

从古生菌、超嗜热菌和真核生物基因组到丙酮丁醇梭菌基因组的基因水平转移，使丙酮丁醇梭菌的代谢功能趋于成形[101]，该菌基因组中的固氮操纵子，在另一固氮古生菌热自养甲烷杆菌（*Methanobacterium thermoautotrophicum*）基因组中也存在；该菌基因组中的芳香族氨基酸生物合成操纵子，在超嗜热菌海栖热袍菌（*Thermotoga maritima*）基因组中也存在。通过对丙酮丁醇梭菌基因组进行分析，发现了与多糖降解有关的一组完整的基因，其中包括纤维素降解基因、木聚糖降解基因、果聚糖降解基因、果胶降解基因和淀粉降解基因。最值得注意的是，在丙酮丁醇梭菌基因组中发现了与有机多聚物胞外水解有关的独特代谢系统基因簇，簇内所有成员都有独特的 ChW 重复疏水序列，其内含有保守的色氨酸密码子，ChW 重复序列位于几个能编码多糖水解酶和蛋白水解酶的可读框上，框内的酶基因很可能与多糖降解和蛋白质水解有关。有趣的是，

在产单核细胞李斯特菌基因组中，也发现此系统中的一个成员存在，此成员具有一个富含亮氨酸区域，研究已表明，这一成员在产单核细胞李斯特菌传染寄主细胞的过程中发挥极其重要的作用[103]。

产气荚膜梭菌基因组

产气荚膜梭菌土壤天然分离株 13 是 A 型菌株，它能引起人类气性坏疽病[104]，对它的基因组测序已经完成，其染色体大小为3 031 430bp 环状 DNA，平均 G + C 含量为 28.6%，推测有 2660 个 ORF[105]。与丙酮丁醇梭菌基因组类似[101]，产气荚膜梭菌基因组中没有可移动遗传元件，其区别是，产气荚膜梭菌基因组中含有 1 个隐秘原噬菌体序列和 7 个完整转座酶基因，而丙酮丁醇梭菌基因组中含有 2 个原噬菌体序列和 3 个转座酶基因，而且，这两种梭状芽孢杆菌基因组没有全面的共线性，它们大部分同源基因在染色体中的位置刚好相反[105]。与芽孢杆菌属中的枯草芽孢杆菌[25]相比，梭状芽孢杆菌属内的产气荚膜梭菌和丙酮丁醇梭菌，都与枯草芽孢杆菌基因组中 61 个与芽孢形成有关基因的核心序列同源，但是，它们缺乏另外 80 多个与芽孢形成有关的基因，其中包括芽孢衣基因和与出芽有关的蛋白基因，这些差异表明，关系紧密的这两个属革兰氏阳性细菌具有完全不同的芽孢形成机制。产气荚膜梭菌基因组和丙酮丁醇梭菌基因组间的另外差异是，后者包含氨基酸生成合成的整套基因，而前者则缺失了许多与氨基酸生物合成有关的基因，因此，它所需要的许多氨基酸很可能是靠寄主细胞提供[101]。

产气荚膜梭菌和丙酮丁醇梭菌两基因组间最大的差异，应该是与致病生化方式有关的基因，因为前者有致病性而后者没有。早先未知的几个与致病有关的基因，在产气荚膜梭菌基因组中鉴定到，其中包括多种溶血素、蛋白酶、透明质酸酶、唾液酸酶、一种内毒素和两种寄主黏附因子——纤维素结合蛋白与胶原质黏附素[105]。与葡萄球菌（*Staphylococci*）和链球菌（*Streptococcui*）等其他革兰氏阳性致病菌不同，产气荚膜梭菌基因组中与致病有关的基因不是在致病基因岛上，而且，大部分毒性因子受 VirR/VirS 和双组分信号传导系统调节。产气荚膜梭菌与其他革兰氏阳性致病菌的另外不同是，后者的毒素及与致病有关的酶，在生长后期或在感染寄主时才产生，而前者的有关基因在生长前期开始表达，这反映了产气荚膜梭菌是通过降解寄主组织而获得氨基酸等营养成分。这点与肺炎链球菌和其他革兰氏阳性致病菌类似，它们的降解酶类执行双重功能，即既降解寄主组织又获得菌体所需要营养[45]。

对艰难梭菌和肉毒梭菌两种致病梭状芽孢杆菌的基因组测序正在进行，这将有利于它们的基因组与丙酮丁醇梭菌基因组的比较，从而能进一步鉴定它们共同的代谢和降解能力。

链球菌

链球菌（*Streptococcus*）是一大群多样的革兰氏阳性球菌，有的在动物体内共生，有的营寄生生活，并能引起动物寄主感病，它们寄生的部位各不相同，有的是众所周知的致病菌，能引起寄主严重和经常的致命疾病，如伤口感染、肺炎、败血症、心内膜炎和猩红热等[106~108]。根据溶血性将链球菌分为两大类：一类是甲型溶血链球菌（α-

hemolytic Streptococci)，包括肺炎链球菌和变异链球菌（*Streptococcus mutans*）等口腔链球菌或草绿色链球菌；另一类是乙型溶血链球菌（又称溶血性链球菌）（*β-hemolytic Streptococci*)，根据其表面糖类抗原再下分为 A-G 族链球菌：A 族链球菌（GAS）是人类致病菌，大多数为化脓链球菌（*Streptococcus pyogenes*)，B 族链球菌（GBS）是人和动物的致病菌，大多数为无乳链球菌（*S. agalactiae*)。已经完成全基因组测序的有肺炎链球菌[45]、化脓链球菌（GAS)[29,31,109]、无乳链球菌（GBS)[22,110]和变异链球菌[111]。

肺炎链球菌基因组

肺炎链球菌是儿童的主要致病菌，常常引起中耳炎和急性呼吸道感染，从而导致全球每年约1 100 000人死亡[107]。肺炎链球菌对青霉素抗性的增加，以及抗多种抗生素菌株的出现，给肺炎链球菌引起传染病的临床治疗带来极大困难。对肺炎链球菌血清型 IV 毒性分离株 TIGR4[45]和非毒性分离株 R6[112]的基因组测序已完成，这为研究肺炎链球菌的致病机制和探究其所致疾病的新治疗方法提供了方便。

分离株 TIGR4 基因组大小为2 160 837bp 环状染色体，平均 G+C 含量为 39.7%，推测有 2236 个 ORF。分离株 R6 的染色体为2 038 615bp 环状 DNA，平均 G+C 含量为 40%，推测有 2043 个 ORF。菌株 TIGR4 和菌株 R6 之间的最大差别是，在分离株 R6 基因组中与荚膜生物合成有关 18kb 区段约有 7kb 发生了缺失。肺炎链球菌基因组中含有较为丰富的插入序列（IS)，例如分离株 TIGR4 基因组中的插入序列（IS）占全基因组 5%[45]，这为与两边序列的同源重组提供了同源区段。有趣的是，在菌株 TIGR4 和菌株 R6 基因组中，都没有鉴定到前噬菌体序列，这与 A 族链球菌（GAS）形成鲜明对比，A 族链球菌（GAS）基因组中，前噬体序列占全基因组 7%～12%，从而造成不同链球菌间的基因组差异和某些菌株毒性的增加。肺炎链球菌还有三个显著特点，一是糖类运输装置的数量很多，二是与寄主组织降解有关的胞外酶系统具有双重作用，三是完全没有三羧酸循环（TCA）代谢途径。首先，肺炎链球菌 30%以上的运输装置转运糖类（见第 7 章)，这可能是生长在富含糖蛋白和胞壁多糖的呼吸道环境中的反映；其次，肺炎链球菌基因组可编码胞外水解酶类和唾液酸苷酶类，这些酶类可能参与寄主组织和多聚物的降解，从而有利于菌体对寄主的侵染，同时，这些降解物又可转运回细菌细胞内，充当基本生物合成的底物，这样肺炎链球菌基因组所编码的胞外酶类执行双重功能。类似的双重功能胞外酶系统，在无乳链球菌[22,110]、金黄色葡萄球菌（*S. aureus*)[27]和产气荚膜梭菌[105]中也鉴定到，而且，在其他革兰氏阳性致病菌中也可能存在；最后，因为肺炎链球菌完全没有 TCA 循环代谢途径，所以不能合成大部分氨基酸前体物质，在其他已完成测序的 A 族链球菌（GAS）和 B 族链球菌（GBS）中，都缺乏 TCA 循环代谢途径。

A 族链球菌

A 族链球菌（GAS）是严格的人类致病菌，能引起咽炎、猩红热、急性风湿热、脓疱病、败血症、坏死性筋膜炎和中毒性休克综合征等多种人类疾病[113]。A 族链球菌（GAS）主要依据 M 蛋白质血清学差异进行菌株分类，M 蛋白质是菌体细胞表面抗噬菌

细胞吞食的蛋白分子，带有特定的 M 血清型，并决定不同菌株的选择性感染类型。化脓链球菌（GAS）三种血清型菌株的基因组测序已经完成，它们是血清型 M1 菌株 SF370，血清型 M3 菌株 MGAS315 和血清型 M18 菌株 MGAS8232。血清型 M1 菌株 SF370 能引起咽炎和侵入性感染[31]，血清型 M3 菌株 MGAS315 常引起死亡率极高的严重传染病[29]，血清型 M18 菌株 MGAS8232 曾在美国引起急性风湿热的爆发流行[109]。

这三种 A 族链球菌（GAS）菌株的基因组大小均为 1.9Mb 左右，平均 G+C 含量为 38%左右，推测有 1752～1889 个 ORF。在这三种 A 族链球菌（GAS）菌株的基因组中，有一段 1.7Mb 核心区域在结构上有共线性，这段核心区域编码链球菌已知的一些毒性因子，如链球菌溶血素 O 和荚膜透明质酸[113]。这三种 A 族链球菌（GAS）菌株基因组的差异，发生在这段核心区域外的原噬菌体序列上，这些原噬菌体序列在三种 A 族链球菌（GAS）菌株基因组中的数量、核苷酸组成和整合位点均不相同。三种 A 族链球菌（GAS）菌株典型的原噬菌体序列大小为 30～50kb，但其数量在不同基因组间有差异，血清型 M3 菌株 MGAS315 基因组中有 6 个原噬菌体序列，占全基因组 12.4%；血清型 M18 菌株 MGAS8232 基因组中有 5 个原噬菌体序列，占全基因组 10.8%；血清型 M1 菌株 SF370 基因组中有 4 个原噬菌体序列，占全基因组 7.0%。这些原噬菌体序列可编码多种毒性因子，如链球菌致热外毒素 K（SpeK），赋予每个菌株独特的毒性特征。因为，原噬菌体序列总是位于不同菌株的同源区段上，而且编码类似的毒性因子，所以，Beres 等[29]认为，不同原噬菌体序列上毒性基因间的同源重组，导致了新的具有更高毒性的 A 族链球菌（GAS）亚克隆菌株群的产生。根据其他相关细菌基因组序列分析，原噬菌体序列对不同 A 族链球菌（GAS）基因组间差异性的产生有很大影响。

B 族链球菌

无乳链球菌为 B 族链球菌（GBS），它首先确认能引起牛乳腺炎，无乳链球菌是典型的妇女肠道和生殖道共生菌，25%～40%健康女性生殖道和胃肠道中存在此细菌。无乳链球菌也是机会致病菌，它能引起孕妇、新生婴儿和易感人群的传染性疾病并能危及生命[114,115]。免疫缺陷通常是成年人感染无乳链球菌所致严重传染病的主要病因，虽然无乳链球菌和其他致病链球菌有共同的毒性决定因素，科学家还是对无乳链球菌基因组进行了测序，以求能鉴定到特有的毒性基因，因为，正是这些毒性基因与独特致病性和基因组进化分离机制密切相关。所有这些问题的解决，将会进一步阐明无乳链球菌如何成为人类的一大主要致病菌。

B 族链球菌（GBS）两个分离株已完成基因组测序，它们是无乳链球菌血清型 III 菌株 NEM316[22]和血清型 V 菌株 2603V/R[110]，前者基因组为2 211 485bp 环状染色体，G+C 含量为 35%，推测有 2082 个 ORF；后者基因组为2 160 267bp 环状染色体，G+C 含量为 35%，推测有 2175 个 ORF，无乳链球菌有多种代谢途径，这反映了它具有适应多种不同寄主环境的能力。首先，与肺炎链球菌类似，在无乳链球菌基因组中，鉴定到一系列依赖磷酸烯醇式丙酮酸的糖特异性磷酸转移酶系统——酶 II 复合物基因，它们编码的酶类可能有很广泛的分解代谢能力；其次，与肺炎链球菌胞外酶双重功能类似，菌株 NEM316 可编码 4 种胞外肽酶、3 种寡肽特异性 ABC 运输装置和多种胞内多

肽酶，这些酶类的协同作用将寄主蛋白质降解，并将降解产物——寡肽转运回菌体细胞作为基本营养成分；最后，与其他链球菌相比，无乳链球菌基因组能编码更多双组分调节系统，这或许反映了它具有更强监控多种寄主环境的能力，以及控制毒性因子表达的能力[110]。

不同B族链球菌（GBS）基因组间的差异是由各种可移动遗传元件引起，其中包括致病基因岛、原噬菌体序列和在无乳链球菌菌株NEM316基因组中鉴定的新型整合质粒。无乳链球菌特有315个基因和主要毒性因子，都分散在基因组多个致病基因岛中。与B族链球菌（GBS）基因组中可移动元件使不同GBS菌株产生多样性类似，A族链球菌（GAS）基因组中可移动遗传元件也使不同GAS菌株获得不同外源基因，从而产生基因组多样性。

变异链球菌

变异链球菌（*Streptococcus mutans*）是口腔微生物菌群中的关键成员，是人类龋齿的罪魁祸首，与格氏链球菌（*S. gordonii*）等其他口腔革兰氏阳性链球菌一样，变异链球菌也与人类细菌性心内膜炎有关[116]。变异链球菌菌株UA159属Bratthall血清型c型，其基因组大小为2 030 936bp环状染色体，平均G+C含量为36.8%，推测有1963个ORF[111]，与肺炎链球菌菌株TIGR4和菌株R6类似，变异链球菌的基因组中也没有原噬菌体序列，但有数量可观的IS元件[45,112]。另外，变异链球菌基因组中含有潜在的接合转座子TnSmu1和40kb的致病基因岛[111]，其中TnSmu1与粪肠球菌基因组中的Tn916极为相似。

变异链球菌能在富含糖类的口腔中生存，这可能预示着它具有比其他任何已完成基因组测序的革兰氏阳性细菌，有更强的代谢多种糖类的能力[111]。与其他革兰氏阳性细菌一样，变异链球菌中大部分糖类是由依赖磷酸烯醇丙酮酸的糖类，经磷酸转移酶系统运输，糖类代谢的最终产物，使口腔环境局部酸化并引起口腔微生物菌群变化，从而更有利于变异链球菌生长，龋齿的形成也由此开始。在变异链球菌中鉴定到多种毒性因子，这些因子包括多种黏附素、蛋白酶、胞外酶和表面蛋白等，新近鉴定的黏附素，包括一种与链球菌胞外基质结合蛋白（matrix-binding protein）类似的蛋白质和一种纤维结合蛋白。在变异链球菌中，还鉴定到几种蛋白酶，它们主要参与寄主细胞结构蛋白的降解，降解产物同时也能作为菌体的营养成分而被转运回菌体细胞，其中的一种蛋白酶与格氏链球菌（*S. gadonii*）的HtpX蛋白酶极为类似[117]，在变异链球菌中还鉴定到许多新细胞表面蛋白，其中包括6种含有C端LPXTG模体的细胞表面蛋白。

其他链球菌基因组

目前，至少有10种其他链球菌正在进行基因组测序或已完成了基因组测序，其中正在进行基因组测序的有，多形链球菌（*S. pleomorphus*）、猪链球菌（*S. suis*）、乳房链球菌（*S. uberis*）、马链球菌（*S. equi*）、嗜热链球菌（*S. thermophilus*）和几种其他口腔链球菌，如血链球菌（*S. anguis*）、格氏链球菌、缓症链球菌（*S. mitis*）和表兄链球菌（*S. sobrinus*）。

乳球菌属

乳酸乳球菌的基因组

乳酸细菌是一类能将糖类转化为乳酸的微生物，其中包括致病性链球菌和其他属的细菌，如乳球菌属和乳杆菌属（*Lactobacillus*）细菌，它们能通过同型乳酸发酵产生乳酸。同型乳酸发酵是乳品加工业的支柱产业，通过其生产的酸乳可用于加工干酪、酸奶酪和其他乳制品[118]。作为乳品加工业的辛勤工作者，乳酸乳球菌（*Lactococcus lactis*）一直受到人们的重视，并开展了大量研究来改善其生长状况和提高乳酸产品的质量[118]。为了对乳酸乳球菌菌株 IL1403 的生理有更全面了解，以求提高更有效的发酵能力，科学家完成了它的基因组测序。

乳酸乳球菌菌株 IL1403 的基因组大小为2 365 589bp，平均 G+C 含量为 35.4%，推测有 2310 个可读框（ORF），基因组中含有相当比例的 IS 元件和原噬菌体序列，它们共占整个基因组的 9.2%，这与已完成基因组测序的化脓链球菌（GAS）和无乳链球菌（GBS）等乳酸细菌基因组中原噬菌体的情形类似[119]，通过基因组测序进行比较分析或将多种乳球菌进行基因组杂交分析，是研究原噬菌体序列引起不同基因组间的差异性必不可少的步骤。

乳酸乳球菌能量代谢是我们最关心的，因为，通过同型乳酸发酵产生大量乳酸[120,121]，厌氧糖酵解是乳酸乳球菌主要的能量产生代谢途径，在基因组中也鉴定到所有与之有关的基因。出乎意料的是乳酸乳球菌基因组也能编码有氧呼吸的酶类，这表明可能有其他产能代谢途径[119]。在乳品加工业中也常利用乳酸乳球菌进行混合发酵，生产乳酸以外的其他发酵产品，通过对乳酸乳球菌的基因组分析，新鉴定到一个染色体编码的丙酮酸氧化酶（PoxL），该酶或许与乳酸乳球菌在不同发酵途径之间的转换有关。

目前，包括嗜酸乳杆菌（*L. acidophilus*）和保加利亚乳杆菌（*L. bulgaricus*）在内的其他乳酸细菌正在进行基因组测序，前者能引起龋齿，同时也能用于生产酸奶等乳制品，后者也能用于生产酸奶酪和干酪等乳制品。

小结

通过对上述所有这些低 G+C 含量的革兰氏阳性细菌进行基因组测序，已经发现了许多新毒性因子和意想不到的代谢途径，以及产生基因组多样性的机制。虽然，这些新毒性因子和独特代谢途径反映了不同细菌对特定环境和不同寄主组织的适应，但是，某些细菌还具有共同的特点，例如，原噬菌体序列和其他可移动遗传因子的插入，对基因组多样性的贡献才刚开始被认识，又如，在许多低 G+C 含量的革兰氏阳性细菌中，许多基因的转录都表现出很强的偏嗜性，即转录方向与复制方向相同（见图 1)。许多致病菌已经进化出多种降解酶类和运输系统，它们协同将寄主组织降解并将降解产物运回到菌体细胞中作为营养物质。继续对多种低 G+C 含量革兰氏阳性细菌的基因组进行比较分析，包括基因组测序和基因组杂交分析，无疑会促进对低 G+C 含量革兰氏阳性细

菌有新的更深入认识，并对基因水平转移在基因组的形成过程中所起的作用有更进一步了解。

（孙 明 译）

参考文献

1. Razin S, Yogev D, Naot Y. Molecular biology and pathogenicity of mycoplasmas. Microbiol Mol Biol Rev 1998; 62:1094–1156.
2. Krause DC. *Mycoplasma pneumoniae* cytadherence: unravelling the tie that binds. Mol Microbiol 1996; 20:247–253.
3. Colman SD, Hu PC, Bott KF. *Mycoplasma pneumoniae* DNA gyrase genes. Mol Microbiol 1990; 4:1129–1134.
4. Su CJ, Baseman JB. Genome size of *Mycoplasma genitalium*. J Bacteriol 1990; 172:4705–4707.
5. Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995; 270:397–403.
6. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 1996; 24:4420–4449.
7. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 2000; 407:757–762.
8. Chambaud I, Heilig R, Ferris S, et al. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res 2001; 29:2145–2153.
9. Hutchison CA, Peterson SN, Gill SR, et al. Global transposon mutagenesis and a minimal *Mycoplasma genome*. Science 1999; 286:2165–2169.
10. Goulet M, Dular R, Tully JG, Billowes G, Kasatiya S. Isolation of *Mycoplasma pneumoniae* from the human urogenital tract. J Clin Microbiol 1995; 33:2823–2825.
11. Baseman JB, Dallo SF, Tully JG, Rose DL. Isolation and characterization of *Mycoplasma genitalium* strains from the human respiratory tract. J Clin Microbiol 1988; 26:2266–2269.
12. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res 1997; 25:701–712.
13. Peterson SN, Fraser CM. The complexity of simplicity. Genome Biol 2001; 2:COMMENT2002.
14. Jensen JS, Hansen HT, Lind K. Isolation of *Mycoplasma genitalium* strains from the male urethra. J Clin Microbiol 1996; 34:286–291.
15. Cassell GH, Waites KB, Watson HL, Crouse DT, Harasawa R. *Ureaplasma urealyticum* intrauterine infection: role in prematurity and disease in newborns. Clin Microbiol Rev 1993; 6:69–87.
16. Smith DG, Russell WC, Ingledew WJ, Thirkell D. Hydrolysis of urea by *Ureaplasma urealyticum* generates a transmembrane potential with resultant ATP synthesis. J Bacteriol 1993; 175:3253–3258.
17. Davidson MK, Lindsey JR, Parker RF, Tully JG, Cassell GH. Differences in virulence for mice among strains of *Mycoplasma pulmonis*. Infect Immun 1988; 56:2156–2162.
18. Shen X, Gumulak J, Yu H, French CT, Zou N, Dylovig K. Gene rearrangements in the *vsa* locus of *Mycoplasma pulmonis*. J Bacteriol 2000; 182:2900–2908.
19. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 1996; 93:10,268–10,273.
20. Vazquez-Boland JA, Kuhn M, Berche P, et al. *Listeria* pathogenesis and molecular virulence determinants. Clin Microbiol Rev 2001; 14:584–640.

21. Cummins AJ, Fielding AK, McLauchlin J. *Listeria ivanovii* infection in a patient with AIDS. J Infect 1994; 28:89–91.
22. Glaser P, Rusniok C, Buchrieser C, et al. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol Microbiol 2002; 45:1499–1513.
23. Gaillard JL, Berche P, Frehel C, Gouin E, Cossart P. Entry of *L. monocytogenes* into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. Cell 1991; 65:1127–1141.
24. Lecuit M, Vandarmael-Pournin S, Lefort J, et al. A transgenic model for listeriosis: role of internalin in crossing the intestinal barrier. Science 2001; 292:1722–1725.
25. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 1997; 390:249–256.
26. Kuroda M, Kuwahara-Arai K, Hiramatsu K. Identification of the up- and downregulated genes in vancomycin-resistant *Staphylococcus aureus* strains Mu3 and Mu50 by cDNA differential hybridization method. Biochem Biophys Res Commun 2000; 269:485–490.
27. Gill SR, et al. Comparative genomics of *Staphylococcus aureus* and *Staphylococcus epidermidis*. 2002; in preparation.
28. Perna NT, Plunkett G 3rd, Burland V, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 2001; 409:529–533.
29. Beres SB, Sylva GL, Barbian KD, et al. Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. Proc Natl Acad Sci USA 2002; 99:10,078–10,083.
30. Navarre WW, Schneewind O. Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol Mol Biol Rev 1999; 63:174–229.
31. Ferretti JJ, McShan WM, Ajdic D, et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 2001; 98:4658–4663.
32. Kuroda M, Ohta T, Uchiyama I, et al. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 2001; 357:1225–1240.
33. Glaser P, Frangeul L, Buchrieser C, et al. Comparative genomics of *Listeria* species. Science 2001; 294:849–852.
34. Herd M, Kocks C. Gene fragments distinguishing an epidemic-associated strain from a virulent prototype strain of *Listeria monocytogenes* belong to a distinct functional subset of genes and partially cross-hybridize with other *Listeria* species. Infect Immun 2001; 69: 3972–3979.
35. He W, Luchansky JB. Construction of the temperature-sensitive vectors pLUCH80 and pLUCH88 for delivery of Tn917::NotI/SmaI and use of these vectors to derive a circular map of *Listeria monocytogenes* Scott A, a serotype 4b isolate. Appl Environ Microbiol 1997; 63: 3480–3487.
36. Richards MJ, Edwards JR, Culver DH, Gaynes RP. Nosocomial infections in combined medical-surgical intensive care units in the United States. Infect Control Hosp Epidemiol 2000; 21:510–515.
37. Huycke MM, Sahm DF, Gilmore MS. Multiple-drug resistant enterococci: the nature of the problem and an agenda for the future. Emerg Infect Dis 1998; 4:239–249.
38. Sahm DF, Kissinger J , Gilmore MS, et al. In vitro susceptibility studies of vancomycin-resistant *Enterococcus faecalis*. Antimicrob Agents Chemother 1989; 33:1588–1591.
39. Bonten MJ, Willems R, Weinstein RA. Vancomycin-resistant enterococci: why are they here, and where do they come from? Lancet Infect Dis 2001; 1:314–325.
40. Paulsen I. Role of mobile elements in the evolution of vancomycin resistant *Enterococcus faecalis* V583. Science 2003; 299:2071–2074.
41. Bensing BA, Manias DA, Dunny GM. Pheromone cCF10 and plasmid pCF10-encoded regula-

tory molecules act post-transcriptionally to activate expression of downstream conjugation functions. Mol Microbiol 1997; 24:285–294.

42. Bruand C, Le Chatelier E, Ehrlich SD, Janniere L. A fourth class of theta-replicating plasmids: the pAM beta 1 family from Gram-positive bacteria. Proc Natl Acad Sci USA 1993; 90:11,668–11,672.
43. de Freire Bastos MC, Tanimoto K, Clewell DB. Regulation of transfer of the *Enterococcus faecalis* pheromone-responding plasmid pAD1: temperature-sensitive transfer mutants and identification of a new regulatory determinant, traD. J Bacteriol 1997; 179:3250–3259.
44. Shankar N, Baghdayan AS, Gilmore MS. Modulation of virulence within a pathogenicity island in vancomycin-resistant *Enterococcus faecalis*. Nature 2002; 417:746–750.
45. Tettelin H, Nelson KE, Paulson IT, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 2001; 293:498–506.
46. Weigel L, Clewell DB, Gill SR, et al. Genetic analysis of a high-level vancomycin-resistant isolate of *Staphylococcus aureus*. Science 2003; 302:1569–1571.
47. Novak R, Henriques B, Charpentier E, Normark S, Tuomanen E. Emergence of vancomycin tolerance in *Streptococcus pneumoniae*. Nature 1999; 399:590–593.
48. CDC. Vancomycin-resistant *Staphylococcus aureus*—Pennsylvania, 2002. Centers for Disease Control and Prevention. MMWR Morb Mortal Wkly Rep 2002; 51:902.
49. Projan SJ, Novick RP. The molecular basis of pathogenicity. In: Crossley KB, Archer GL (eds). The Staphylococci in Human Disease. New York: Churchill Livingstone, 1997, pp. 55–81.
50. Steinberg JP, Clark CC, Hackman BO. Nosocomial and community-acquired *Staphylococcus aureus* bacteremias from 1980 to 1993: impact of intravascular devices and methicillin resistance. Clin Infect Dis 1996; 23:255–259.
51. Thylefors JD, Harbarth S, Pittet D. Increasing bacteremia due to coagulase-negative staphylococci: fiction or reality? Infect Control Hosp Epidemiol 1998; 19:581–589.
52. Rupp ME, Archer GL. Coagulase-negative staphylococci: pathogens associated with medical progress. Clin Infect Dis 1994; 19:231–243; quiz 244–245.
53. Srinivasan A, Dick JD, Perl TM. Vancomycin resistance in staphylococci. Clin Microbiol Rev 2002; 15:430–438.
54. Groom AV, Wolsey DH, Naimi TS, et al. Community-acquired methicillin-resistant *Staphylococcus aureus* in a rural American Indian community. JAMA 2001; 286:1201–1205.
55. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J Clin Microbiol 2000; 38:1008–1015.
56. Naimi TS, LeDell KH, Boxrud DJ, et al. Epidemiology and clonality of community-acquired methicillin-resistant *Staphylococcus aureus* in Minnesota, 1996–1998. Clin Infect Dis 2001; 33:990–996.
57. Iandolo JJ, Worrell V, Groicher KH, et al. Comparative analysis of the genomes of the temperate bacteriophages phi 11, phi 12 and phi 13 of *Staphylococcus aureus* 8325. Gene 2002; 289: 109–118.
58. Baba T, Takeuchi F, Kuroda M, et al. Genome and virulence determinants of high virulence community-acquired MRSA. Lancet 2002; 359:1819–1827.
59. Novick RP, Schlievert P, Ruzin A. Pathogenicity and resistance islands of staphylococci. Microbes Infect 2001; 3:585–594.
60. Lindsay JA, Ruzin A, Ross HF, Kurepina N, Novick RP. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. Mol Microbiol 1998; 29:527–543.
61. Firth N, Apisiridej S, Berg T, et al. Replication of staphylococcal multiresistance plasmids. J Bacteriol 2000; 182:2170–2178.

62. Hiramatsu K, Cui L, Kuroda M, Ito T. The emergence and evolution of methicillin-resistant *Staphylococcus aureus*. Trends Microbiol 2001; 9:486–493.
63. Ito T, Katayama Y, Asada K, et al. Structural comparison of three types of staphylococcal cassette chromosome mec integrated in the chromosome in methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother 2001; 45:1323–1336.
64. Cui L, Murakami H, Kuwahara-Arai F, Hanaki H, Hiramatsu K. Contribution of a thickened cell wall and its glutamine nonamidated component to the vancomycin resistance expressed by *Staphylococcus aureus* Mu50. Antimicrob Agents Chemother 2000; 44:2276–2285.
65. Heilmann C, et al. Characterization of the 113 kDa giant Staphylococcal surface protein (gssp) from *Staphylococcus aureus* involved in adherence to endothelial cells. International Symposium on Staphylococci and Staphylococcal Infections—Abstracts 2002; 2002:151.
66. Clarke SR, et al. Components of *Staphylococcus aureus* expressed during human infection. General Society for Microbiology—2002 Abstracts, 2002:5.
67. Peng HL, Novick RP, Kreiswirth B, Kornblum J, Schlievert P. Cloning, characterization, and sequencing of an accessory gene regulator (*agr*) in *Staphylococcus aureus*. J Bacteriol 1988; 170:4365–4372.
68. Cheung AL, Projan SJ. Cloning and sequencing of sarA of *Staphylococcus aureus*, a gene required for the expression of agr. J Bacteriol 1994; 176:4168–4172.
69. Cheung AL, Yeaman MR, Sullam PM, Witt MD, Bayer AS. Role of the sar locus of *Staphylococcus aureus* in induction of endocarditis in rabbits. Infect Immun 1994; 62:1719–1725.
70. Alonso JC, Luder G, Strege AC, et al. The complete nucleotide sequence and functional organization of *Bacillus subtilis* bacteriophage SPP1. Gene 1997; 204:201–212.
71. Herron LL, Chakravarty R, Dwan C, et al. Genome sequence survey identifies unique sequences and key virulence genes with unusual rates of amino acid substitution in bovine *Staphylococcus aureus*. Infect Immun 2002; 70:3978–3981.
72. Priest FG. Systematics and ecology of *Bacillus*. In: Sonenshein AL, Hoch JA, Losick R (eds). *Bacillus subtilis* and other Gram-Positive Bacteria. Washington, DC: American Society for Microbiology, 1993.
73. Stragier P, Losick R. Molecular genetics of sporulation in *Bacillus subtilis*. Annu Rev Genet 1996; 30:297–241.
74. Shapiro L, Losick R. Dynamic spatial regulation in the bacterial cell. Cell 2000; 100:89–98.
75. Harwood CR. *Bacillus subtilis* and its relatives: molecular biological and industrial workhorses. Trends Biotechnol 1992; 10:247–256.
76. Read TD, Salzberg SL, Pop M, et al. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 2002; 296:2028–2033.
77. Azevedo V, Alvarez E, Zumstein E, et al. An ordered collection of *Bacillus subtilis* DNA segments cloned in yeast artificial chromosomes. Proc Natl Acad Sci USA 1993; 90:6047–6051.
78. Sorokin A, Lapidus A, Capuano V, et al. A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. Genome Res 1996; 6:448–453.
79. Soppa J. Prokaryotic structural maintenance of chromosomes (SMC) proteins: distribution, phylogeny, and comparison with MukBs and additional prokaryotic and eukaryotic coiled-coil proteins. Gene 2001; 278:253–264.
80. Hirano M, Hirano T. Hinge-mediated dimerization of SMC protein is essential for its dynamic interaction with DNA. EMBO J 2002; 21:5733–5744.
81. Horikoshi K. Alkaliphiles: some applications of their products for biotechnology. Microbiol Mol Biol Rev 1999; 63:735–750, table of contents.
82. Takami H, Nakasone K, Takaki Y, et al. Complete genome sequence of the alkaliphilic bacte-

rium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 2000 28:4317–4331.

83. Takami H, Han CG, Takaki Y, Ontsuno E. Identification and distribution of new insertion sequences in the genome of alkaliphilic *Bacillus halodurans* C-125. J Bacteriol 2001; 183: 4345–4356.
84. Pasteur L, Chamberland, Roux. Summary report of the experiments conducted at Pouilly-le-Fort, near Melun, on the anthrax vaccination. 1881 [classical article]. Yale J Biol Med 2002; 75:59–62.
85. Okinaka R, Cloud K, Hampton O, et al. Sequence, assembly and analysis of pX01 and pX02. J Appl Microbiol 1999; 87:261–262.
86. Okinaka RT, et al. Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. J Bacteriol 1999; 181:6509–6515.
87. Read TD, Peterson SN, Tourasse N, et al. The complete genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 2003; 423:81–86.
88. Mignot T, Mock M, Robichon D, et al. The incompatibility between the PlcR- and AtxA-controlled regulons may have selected a nonsense mutation in *Bacillus anthracis*. Mol Microbiol 2001; 42:1189–1198.
89. Agaisse H, Gominet M, Okstad OA, Kolsto AB, Lereclus D. PlcR is a pleiotropic regulator of extracellular virulence factor gene expression in *Bacillus thuringiensis*. Mol Microbiol 1999; 32:1043–1053.
90. Helgason E, Okstad O A, Caugant DA, et al. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* — one species on the basis of genetic evidence. Appl Environ Microbiol 2000; 66: 2627–2630.
91. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 2002; 30:2478–2483.
92. Stackebrandt E, Kramer I, Swiderski J, Heppe H. Phylogenetic basis for a taxonomic dissection of the genus *Clostridium*. FEMS Immunol Med Microbiol 1999; 24:253–258.
93. Keis S, Bennett CF, Ward VK, Jones DT. Taxonomy and phylogeny of industrial solvent-producing clostridia. Int J Syst Bacteriol 1995; 45:693–705.
94. Durre P. New insights and novel developments in clostridial acetone/butanol/isopropanol fermentation. Appl Microbiol Biotechnol 1998; 49:639–648.
95. Woods DR. The genetic engineering of microbial solvent production. Trends Biotechnol 1995; 13:259–264.
96. Mitchell WJ. Physiology of carbohydrate to solvent conversion by clostridia. Adv Microb Physiol 1998; 39:31–130.
97. Ravagnani A, Jennert KC, Steiner E, et al. Spo0A directly controls the switch from acid to solvent production in solvent-forming clostridia. Mol Microbiol 2000; 37:1172–1185.
98. Green EM, Boynton ZL, Harris LM, et al. Genetic manipulation of acid formation pathways by gene inactivation in *Clostridium acetobutylicum* ATCC 824. Microbiology 1996; 142(Pt 8): 2079–2086.
99. Hatheway CL. Toxigenic clostridia. Clin Microbiol Rev 1990; 3:66–98.
100. Rood JI. Virulence genes of *Clostridium perfringens*. Annu Rev Microbiol 1998; 52:333–360.
101. Nolling J, Breton G, Omelchenko MV, et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol 2001; 183:4823–4838.
102. Cornillot E, Nair RV, Papoutsakis ET, Soucaille P. The genes for butanol and acetone formation in *Clostridium acetobutylicum* ATCC 824 reside on a large plasmid whose loss leads to degeneration of the strain. J Bacteriol 1997; 179:5442–5447.
103. Marino M, Braun L, Cossart P, Ghosh P. Structure of the InlB leucine-rich repeats, a domain

that triggers host cell invasion by the bacterial pathogen *L. monocytogenes*. Mol Cell 1999; 4:1063–1072.

104. Mahony DE, Moore TI. Stable L-forms of *Clostridium perfringens* and their growth on glass surfaces. Can J Microbiol 1976; 22:953–959.
105. Shimizu T, et al. Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. Proc Natl Acad Sci USA 2002; 99:996–1001.
106. Greenwood B. The epidemiology of pneumococcal infection in children in the developing world. Philos Trans R Soc Lond B Biol Sci 1999; 354:777–785.
107. Klein DL, Eskola J. Development and testing of *Streptococcus pneumoniae* conjugate vaccines. Clin Microbiol Infect 1999; 5(Suppl 4):S17–S28.
108. Musher DM. Infections caused by *Streptococcus pneumoniae*: clinical spectrum, pathogenesis, immunity, and treatment. Clin Infect Dis 1992; 14:801–807.
109. Smoot JC, Barbian KD, Van Gompel JJ, et al. Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. Proc Natl Acad Sci USA 2002; 99:4668–4673.
110. Tettelin H, Masignani V, Cieslewicz MJ, et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci USA 2002; 99:12,391–12,396.
111. Ajdic D, McShan WM, McLaughlin RE, et al. Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. Proc Natl Acad Sci USA 2002; 99:14,434–14,439.
112. Hoskins J, Alborn WE Jr, Arnold J, et al. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 2001; 183:5709–5717.
113. Cunningham MW. Pathogenesis of group A streptococcal infections. Clin Microbiol Rev 2000; 13:470–511.
114. Schuchat A, Deaver-Robinson K, Plikaytis BD, et al. Multistate case-control study of maternal risk factors for neonatal group B streptococcal disease. The Active Surveillance Study Group. Pediatr Infect Dis J 1994; 13:623–629.
115. Schuchat A, Wenger JD. Epidemiology of group B streptococcal disease. Risk factors, prevention strategies, and vaccine development. Epidemiol Rev 1994; 16:374–402.
116. Herzberg MC. In: Stevens DL, Kaplan EL (eds). Streptococcal Infections. New York: Oxford University Press, 2000, pp. 333–370.
117. Vickerman MM, Mathu NM, Minick PE, Edwards CA. Initial characterization of the *Streptococcus gordonii* htpX gene. Oral Microbiol Immunol 2002; 17:22–31.
118. Kleerebezem M, Boels IC, Groot MN, et al. Metabolic engineering of *Lactococcus lactis*: the impact of genomics and metabolic modelling. J Biotechnol 2002; 98:199–213.
119. Bolotin A, Wincker P, Mauger S, et al. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp lactis IL1403. Genome Res 2001; 11:731–753.
120. Cocaign-Bousquet M, Even S, Lindley ND, Loubiere P. Anaerobic sugar catabolism in *Lactococcus lactis*: genetic regulation and enzyme control over pathway flux. Appl Microbiol Biotechnol 2002; 60:24–32.
121. Tanaka K, Komiyama A, Sonomoto K, et al. Two different pathways for D-xylose metabolism and the effect of xylose concentration on the yield coefficient of L-lactate in mixed-acid fermentation by the lactic acid bacterium *Lactococcus lactis* IO-1. Appl Microbiol Biotechnol 2002; 60:160–167.

19

放线菌（G$^+$，高 G +C 含量）基因组学

Stephen D. Bentley, Roland Brosch, Stephen V. Gordon, David A. Hopwood, and Stewart T. Cole

放线菌引言

Woese 及其同事将生物划分为三域：古生菌、细菌和真核生物[1]。细菌又分为 15 个纲，放线菌纲是其中之一。放线菌的共同特点是革兰氏阳性，G + C 含量较高；其进一步分类依据为：化学特征（肽聚糖、脂等化合物的化学组分）、DNA 碱基组成（G + C 大于 50%）和 16S rRNA 序列。根据 16S rRNA 序列相似性，Stackebrandt 等将此类

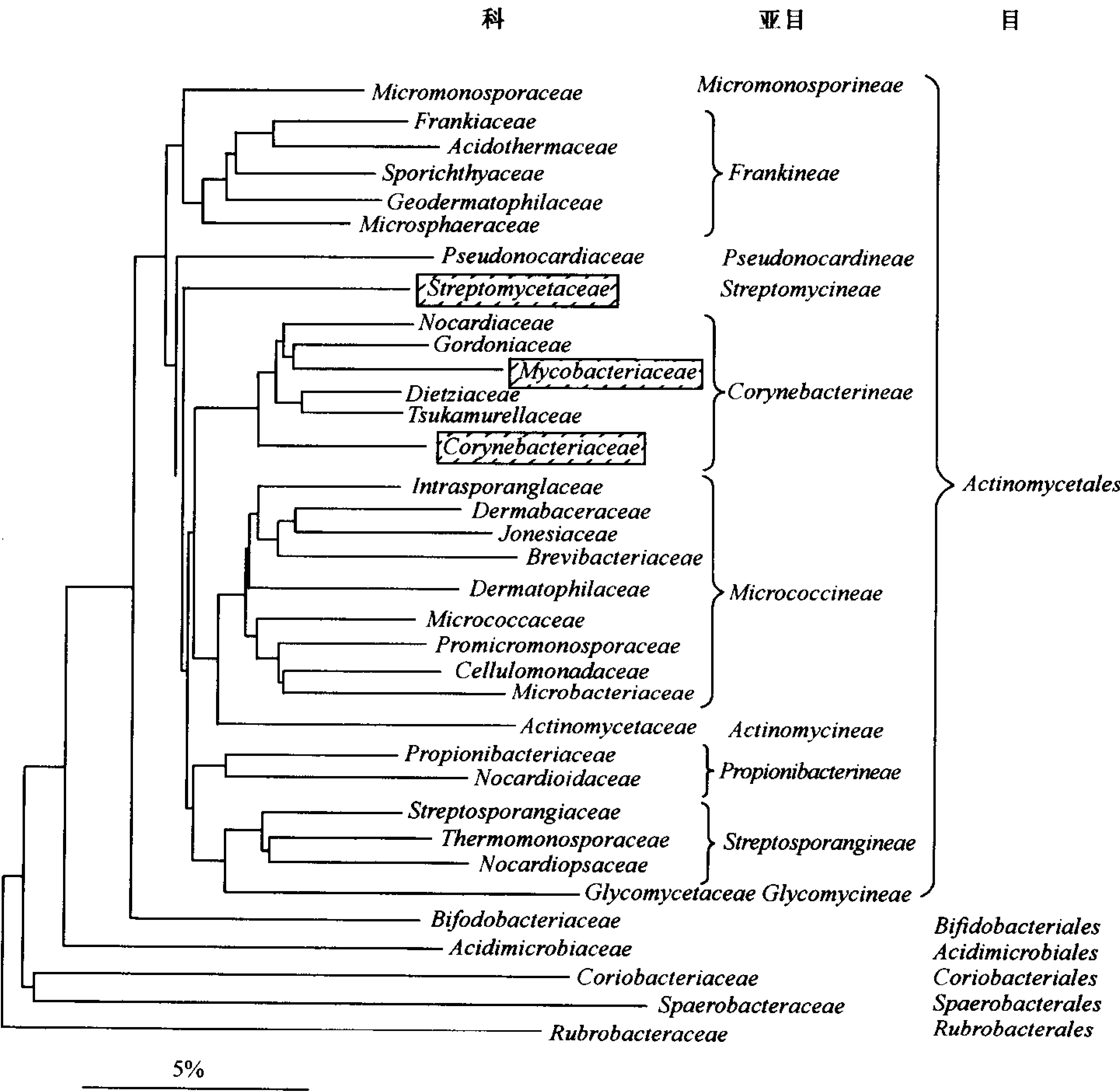


图 1　某些放线菌的种系进化关系（根据文献［2]）。比例尺代表每 100 个 16S rDNA/rRNA 核苷酸序列有 5 个核苷酸的区别；方框表示该科的某成员基因组已被测序（截至本书成稿时）。

细菌定名为放线菌纲[2]，纲下面为放线菌目（*Actinomycetales*），进一步划分为 10 个亚目，本章只讨论其中的几个代表科（见图 1）。

放线菌的种间差异很大，链霉菌（*Streptomyces*）的一些种具有包括产孢过程的复杂生活史，有些致病分枝杆菌在经过开始的生长期后，能进入长久的非繁殖状态；它具有从腐生型到严格致病型多种生活方式；其菌落形态和色素丰富多彩；对放线菌现有基因组分析发现，它的生物多样性表明它们在人类医学、兽医学、生物技术和生态学等多种领域的重要性。

已完全测序的放线菌基因组有 7 种，如表 1 所列，种间特征差别极大，基因组小的仅 2.5 Mb，大的有 8.7 Mb，基因数从 1600 到 7800 个以上，rRNA 操纵子数从 1 个（如生长缓慢的分枝杆菌）到 7 个（如生长迅速的链霉菌和棒杆菌），G + C 含量从棒杆菌的 53.5%到链霉菌的 72.1%。麻风分枝杆菌（*Mycobacterium leprae*）的基因衰变十分普遍[3]，因此其基因密度最低，仅为其他种的一半（每 2 kb 序列仅 1 个基因）（表 1）。

表 1　基因组特征比较

特征	白喉杆菌 (*C. diphtheriae*) NCTC13139	谷氨酸棒杆菌 (*C. glutamicum*) ATCC13032	麻风分枝杆菌 (*M. leprae*) TN	牛分枝杆菌 (*M. bovis*) AF2122/97	结核分枝杆菌 (*M. tuberculosis*) H37Rv	结核分枝杆菌 (*M. tuberculosis*) CDC1551	天蓝色链霉菌 (*S. coelicolor*) A3(2)
基因组大小/bp	2 488 635	3 309 401	3 268 203	4 345 492	4 411 532	4 403 836	8 667 507
G+C 百分含量/%	53.50	54.72	57.79	65.63	65.61	65.60	72.12
蛋白质编码/%	87.9	87.3	49.5	90.8	90.8	92.7	88.9
编码蛋白质的基因数/个	2320	3099	1605	3953	3994	4250	7825
假基因	45	NA	1116	23	6[a]	NA	55
基因密度/(bp/gene)	1073	1067	20.7	1108	1104	1036	1107
基因平均长度/bp	962	933	1011	1003	1004	960	991
未知基因平均长度/bp	NA	NA	338	653	584	NA	NA
rRNA 操纵子/个	5	6	1	1	1	1	6
rRNA	54	60	45	45	45	45	63
其他稳定 RNA	NA	2	2	2	2	2	3

注：NA，无数据；[a]不包含 IS 插入因子。

最近刚完成基因组测序的两个棒杆菌是：白喉杆菌（*Corynebacterium diphtheriae*）和谷氨酸棒杆菌（*C. glutamicum*）。白喉杆菌（*C. diphtheriae*）产毒菌株导致急性传染性呼吸道疾病，期望其基因组测序有助于该病的预防与治疗。而谷氨酸棒杆菌（*C. glutamicum*）则是被广泛利用的工业微生物，用于氨基酸的发酵生产，因此，其基因组研究的重点是菌株改造，以构建高产菌。棒杆菌基因组测序后，将有更多关于基因组分析的文献问世。本章的重点是已发表的放线菌基因组。

结核分枝杆菌基因组序列

引言

结核分枝杆菌（*Mycobacterium tuberculosis*）H37Rv 是人类肺结核致病菌的典型菌株[4]，在其复合群中各菌株亲缘关系密切，H37Rv 是其中的一个代表菌株，也是第一个测序的放线菌基因组。这个复合群中的其他成员有：非洲分枝杆菌（*M. africanum*）（非洲一些地方的肺结核致病菌）、牛分枝杆菌（*M. bovis*）（小牛肺结核致病菌）、*M. bovis* BCG（肺结核活疫苗）、*M. canettii*（罕发性人类致病菌，菌落表面光滑）和田鼠分枝杆菌（*M. microti*）（田鼠和鼱鼠的致病菌，对人无毒）。该复合群紧密相关，早期遗传学分析表明各菌株在 DNA 水平上极其相似[5]，因此，有人推断结核病直到15 000年前才在人类身上出现[6]。

结核分枝杆菌的基因组学

结核分枝杆菌 H37Rv 基因组大小为 4.41 Mb，较早的生物信息分析预测 H37Rv 有 3924 个蛋白质编码基因，后来运用较低阈值和更新算法，表明有近 4000 个基因[7]。CDC1551[8]是一个新临床分离株，在美国某一乡村传播率极高，导致广泛的皮试阳性反应[9]，其基因组比 H37Rv 略小(4 403 836 bp)（表 1），却极其相似（99.94%）。自动搜寻预测 CDC1511 有 4250 个基因，因为有些估计的基因偏短（约为 30 个密码子），实际值可能略低。

重复 DNA 与可转移遗传元件

基因复制和重复 DNA 元件是影响染色体结构的主要因素，结核分枝杆菌 H37Rv 有 51%以上的基因是由于基因重复或结构域混组（domain-shuffling）造成的[11]，其基因组的 3.4%由插入元件（IS）和原噬菌体（phiRv1、phiRv2）组成。有 52 个位点带有 IS 插入元件，它们属于已知的 IS*3*、IS*5*、IS*21*、IS*30*、IS*110*、IS*256* 和 ISL*3* 家族，以及一个有 6 个成员的新家族，IS*1535*，可能是通过移码产生重组酶[11]。在结核分枝杆菌 H37Rv 基因组中，还发现一个具有 7 个拷贝的新重复序列 REP13E12 家族，它含有噬菌体 phiRV1 附着位点。有趣的是，与结核分枝杆菌 H37Rv 比较，在菌株 Erdman 和 CDC1551 中，phiRv1 整合在 REP13E12 不同的拷贝上[4,8,11,12]。

IS*6110* 属 IS*3* 家族，在结核分枝杆菌中它是出现频率最高的 IS 元件，因此，在基因组结构中起重要作用。结核分枝杆菌 H37Rv 有 16 份 IS*6110* 拷贝，而 CDC1551 仅有 4 份拷贝，IS*6110* 频繁转座，进而产生限制性片段长度多样性，使它成为流行病学的一个有用工具[13]。转座插入不仅可以导致基因失活，还可以通过插入元件相邻拷贝间遗传物质的丢失，而导致大量基因缺失[14]，例如，在 H37Rv 中，RvD2-RvD5 导致了 9 个以上基因的缺失[15]，位点 RvD2 是导致菌株间极大多样性的一个关键[16]。

结核分枝杆菌基因组学与生物学

基因组序列为更深入认识结核杆菌的生物学提供了宝贵的新信息，例如，8%的基

因组序列与脂肪代谢有关系，可见脂肪代谢在结核杆菌生活史中的重要性。早已众所周知，结核分枝杆菌有一层富含多种脂类化合物的蜡质层[17]，仍然惊奇的是，通过基因组分析发现如此众多的基因和蛋白质参与脂肪代谢。不错，在结核分枝杆菌的寄生环境中，脂肪和固醇作为代谢底物远比碳水化合物丰富。该菌具有典型的β氧化途径用作脂肪分解代谢，该途径由多功能脂肪酸降解蛋白（FadA/FadB）所催化，它同时还有约100种酶可能用于其他脂类氧化途径，用来代谢寄主细胞膜降解时产生的外源脂类物质[4]，由此产生的乙酰辅酶A，可用于微生物细胞壁的合成或三羧酸循环，和用于乙醛酸支路或其他代谢循环。

对前所未知或含糊不清的一些代谢细节，基因组分析提供了新答案，例如，基因组分析发现了一整套合成氨基酸或维生素的基因，这与结核杆菌可在人工合成培养基中生长的事实相一致。与此形成对比的是，其他一些胞内寄生菌则直接从寄主中摄取这些合成途径的产物，因此，这一套基因要么没有，要么只存在于其他途径的某一分支中。尽管这个发现与结核分枝杆菌只是在较晚时期才成为人类病原菌的假说[6]一致，但该发现也极可能表明，在吞噬胞中这些代谢产物缺乏，形成了维持这些基因的一个选择压力。在麻风病杆菌中也存在类似倾向，这将在另一节中讨论[3]。

除脂肪降解外，许多其他物质的分解也可以产生能量，如碳水化合物、醇类、酮类和碳氢化合物。基因组分析也表明，结核分枝杆菌还具有糖酵解和磷酸戊糖途径和200多种可以代谢其他碳化合物的氧化还原酶、氧化酶和脱氢酶。特别指出的是，含细胞色素P450有20个成员的单氧化酶体系，能把氧引入分解或合成代谢中的有机分子中[18,19]；这类酶通常只在降解有机物的土壤微生物中存在。与其他分枝杆菌一样，该体系在结核分枝杆菌中的存在表明，结核分枝杆菌在变成严格病原菌之前，曾栖息过与土壤微生物类似的生态环境。此外，与许多真菌类似，结核分枝杆菌还有在固醇生物合成中起作用的P450酶系[20,21]。

ATP是在有氧生长条件下经氧化磷酸化途径合成的，其中还原态NADH作为电子供体被NADH氧化酶所氧化，释放的电子经电子传递链（包括醌细胞色素*b*还原酶复合体和细胞色素*c*氧化酶）最后转移给分子氧。然而，基因组分析表明，如果有相应的电子受体，结核分枝杆菌也能利用NADH或其他电子供体进行无氧呼吸，可作为电子受体的包括硝酸、亚硝酸和延胡索酸[4]。而突变研究表明，硝酸还原酶在分枝杆菌致病毒力方面具有重要性[22]，在溃疡和肉芽瘤的低氧环境中，这套无氧代谢酶系对结核分枝杆菌在侵染过程中的生长起极为重要的作用。

结核分枝杆菌基因组计划最重要的发现之一，是找出一些前所未知或知之甚少的大基因家族，其中最为重要的是PE（ProGlu）和PPE（ProProGlu）基因家族，分别有100和67个成员[4,23]，每个都含有一个保守N端序列（分别约为110和180个氨基酸），其中第8，9个或第8，9，10个氨基酸是特征性的PE或PPE模体（motif），它们的C端都为高度重复序列。

有特别意义的是属PGRS（polymorphic G + C-rich sequence）亚类的PE蛋白质和属MPTR（major polymorphic tandem repeat）亚类的PPE蛋白质，前者的氨基酸组成几乎半数为甘氨酸，并以模体AsnGlyGlyAlaGlyGlyAla的串联重复形式存在，而后者则含有多次重复的模体AsnXGlyXGlyAsnXGly。近来有些关于这类蛋白质的研究进展，它们有

的属细胞表面蛋白[24~27]。重复序列数的增减使相应蛋白质之间差异极大，因此，这类蛋白质的编码基因是种间和种内多样性的主要来源，有证据表明，它们是各种各样的表面抗原[24,27,28]，但其具体生物学作用却不甚明了，有些与致病性有关[29,30]，有的作为黏附物而影响寄主的吞噬作用。这些前所未知的PE和PPE蛋白质的发现证明了基因组学在研究中的重要性。

牛分枝杆菌的基因组学

引言

牛分枝杆菌（*Mycobaterium bovis*）与结核分枝杆菌（*M. tuberculosis*）复合群在DNA水平上非常相近（大于99.9%），但二者的寄主范围、生理特征和毒性方面却存在区别。结核分枝杆菌几乎不导致家畜或野生动物病害，牛分枝杆菌却是多种动物的有效致病菌；在以甘油为唯一碳源的培养基中，丙酮酸盐是牛分枝杆菌是必需的，结核分枝杆菌却不需要；二者的区别还表现在细胞壁的脂肪组成上，牛分枝杆菌具有含酚的糖脂，结核分枝杆菌却没有。当然，这种表型的区别是由基因组所决定，对基因组分析，特别是对种间相异区域的分析，得以从遗传学角度对牛分枝杆菌的生物学特征勾画出一个轮廓。

牛分枝杆菌与结核分枝杆菌的比较

在牛分枝杆菌基因组序列面世之前，结核分枝杆菌（*M. tuberculosis*）复合群的比较基因组学，采用的是微阵列（microarray）和宏阵列（macroarray）通过杂交的方法，结果发现它们的基因组有由于缺失导致的一些差异区段（region of difference，RD，图2），RD的大小从1～12.7 kb不等[31,32]。尽管这些缺失在进化中的意义尚不清楚，但它们肯定对细菌的致病毒力和一系列代谢活动有影响；可能代表寄主适应性突变，但也许是对多余重复的一种切除性修复。

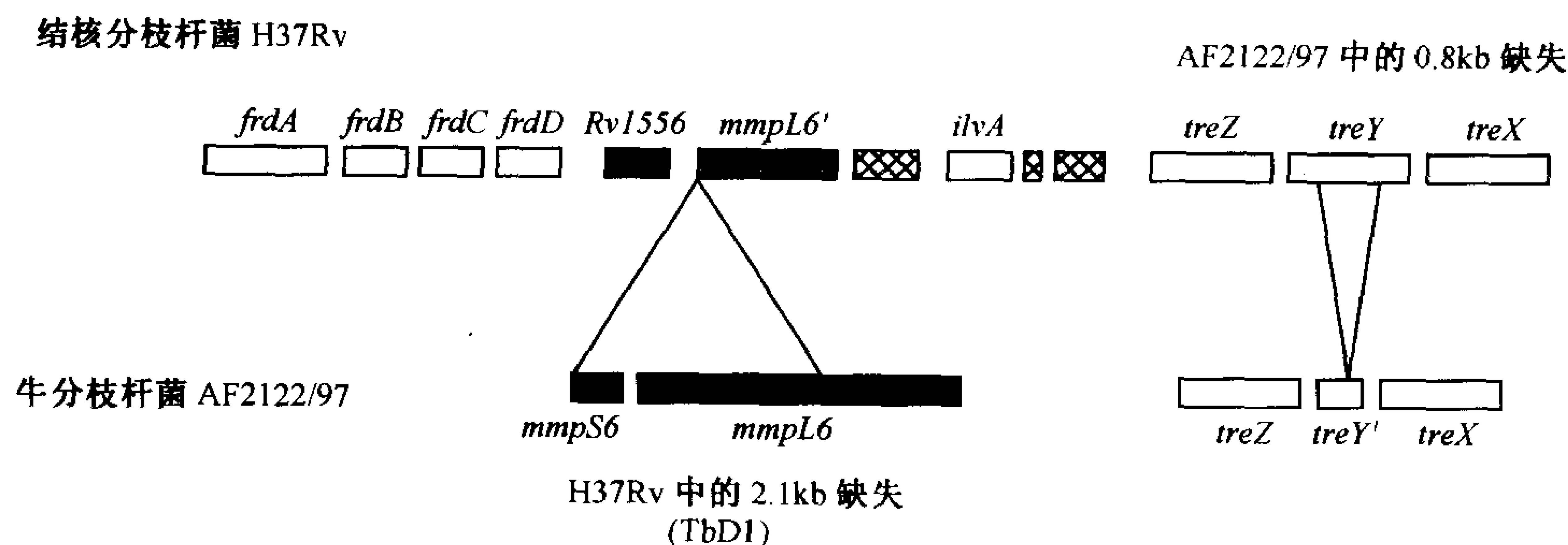


图2 结核分枝杆菌基因组缺失的例子。由于缺失，结核分枝杆菌丢失了大部分*mmpS6*基因和部分*mmpL6*，而这两个基因在牛分枝杆菌中完整无缺（42）；相反，与结核分枝杆菌相比，牛分枝杆菌的*treY*基因缺失。

RD5位点的缺失导致了3个磷脂酶C的丢失，这些酶在李斯特氏菌（*Listeria*）和

梭状芽孢杆菌（*Clostridium*）中是已知的毒力因子[33]。但牛分枝杆菌却保留着第 4 个完整磷脂酶基因 *plcD*，它可能会补偿丢失的其他磷脂酶基因，RD7 位点包含一个 *mce* 操纵子，基因组序列表明结核分枝杆菌含有 4 个 *mce* 操纵子，编码由 24 个蛋白质组成的蛋白家族[4]。因此，随 RD7 丢失的一个 *mce* 操纵子可能由其他剩余的来补偿。Arruda 等认为，RD7 中的 *mce* 操纵子可能编码分枝杆菌侵染素（invasin）[34]。

显然，基因缺失是影响牛分枝杆菌基因组组成的重要因素，然而，从影响物种的生物学角度看，单核苷酸多型性（SNP）具有同等重要性。研究表明，通过 *pncA* 基因突变，牛分枝杆菌对吡嗪酰胺（pyrazinamide）的可遗传抗性就是由于 SNP 引起的[35]，主要 sigma 因子中，一个碱基的改变足够可以减弱牛分枝杆菌的毒性[36]。值得注意的是，与结核分枝杆菌相比，牛分枝杆菌的 *sigM* 有一个阅读框移码，这种变化可能影响特定环境下基因调控元（regulon）的转录，因此，可能与寄主适应性突变有关。

抗原变异

与其他结核杆菌的明显区别是，牛分枝杆菌含有两个超常表达的优势血清（serodominant）抗原，Mpb70 和 Mpb83，其中 Mpb70 占培养体滤液蛋白总量的 10%[37]。修饰表面抗原是细菌对免疫调节的一种策略，这些抗原蛋白也可能直接与细菌的毒力和致病性有关。在牛分枝杆菌中，受基因组缺失影响的另一类抗原是 ESAT-6 家族，它是最早报道为结核分枝杆菌分泌型高效 T 细胞抗原[38]。蛋白组分析表明，ESAT-6 属于由 22 个蛋白质（包括 CFP-10 和 CFP-7）组成的 T 细胞抗原家族[10]，研究还表明 ESAT-6 与 CFP-10 有相互作用，表明该家族其他蛋白质也可能以混合配对（mix-and-match）的方式发挥作用[39]。但是，牛分枝杆菌却缺乏由 RD5 和 RD8 位点编码该家族蛋白质的 4 个成员，这种缺失会影响其他成员蛋白质的功用（假定这些蛋白是成对作用），却很难预测这种缺失的具体影响。

细胞壁的变异

致病细菌细胞壁成分的蛋白质序列存在差异，表明选择压对细胞壁结构的作用[40,41]，表面蛋白的改变影响抗原的变异、配体-受体的互作以及寄主与病原菌的对话。因此，在结核分枝杆菌复合群中，经常变异的基因多为编码膜蛋白、分泌蛋白或转运蛋白的基因。

最引人注目的例子是 PE-PGRS 和 PPE 蛋白家族的序列变异，尽管这类蛋白质的功能开始不为人知，但现在已有相当多的证据表明，它们中至少有一些是表面蛋白，并有免疫调节作用[24,26,27]。基因组比较分析表明，它们的编码基因也是牛分枝杆菌和结核分枝杆菌之间差异最大的，可列举两例，PPE-MPTR 的 Rv1917c 蛋白在两个菌之间的差异极为显著，如图 3A 所示，牛分枝杆菌的该基因有 3 个分散的插入片段，事实上，Rv1917c 蛋白最开始描述为 *katG* 基因上游的可变区段，因此，用于染色体的微卫星分型（minisatellite typing）。图 3B 则显示，与结核分枝杆菌相比，牛分枝杆菌 Rv3508 蛋白（属于 PE-PGRS）的某些区段具有独特的序列，表明这类蛋白质可以承受很广泛的序列多型性而不丧失其功能，是选择压力发挥作用的理想底物。

仅仅存在于牛分枝杆菌基因组序列中，而结核分枝杆菌大多数菌系缺乏的一个区段

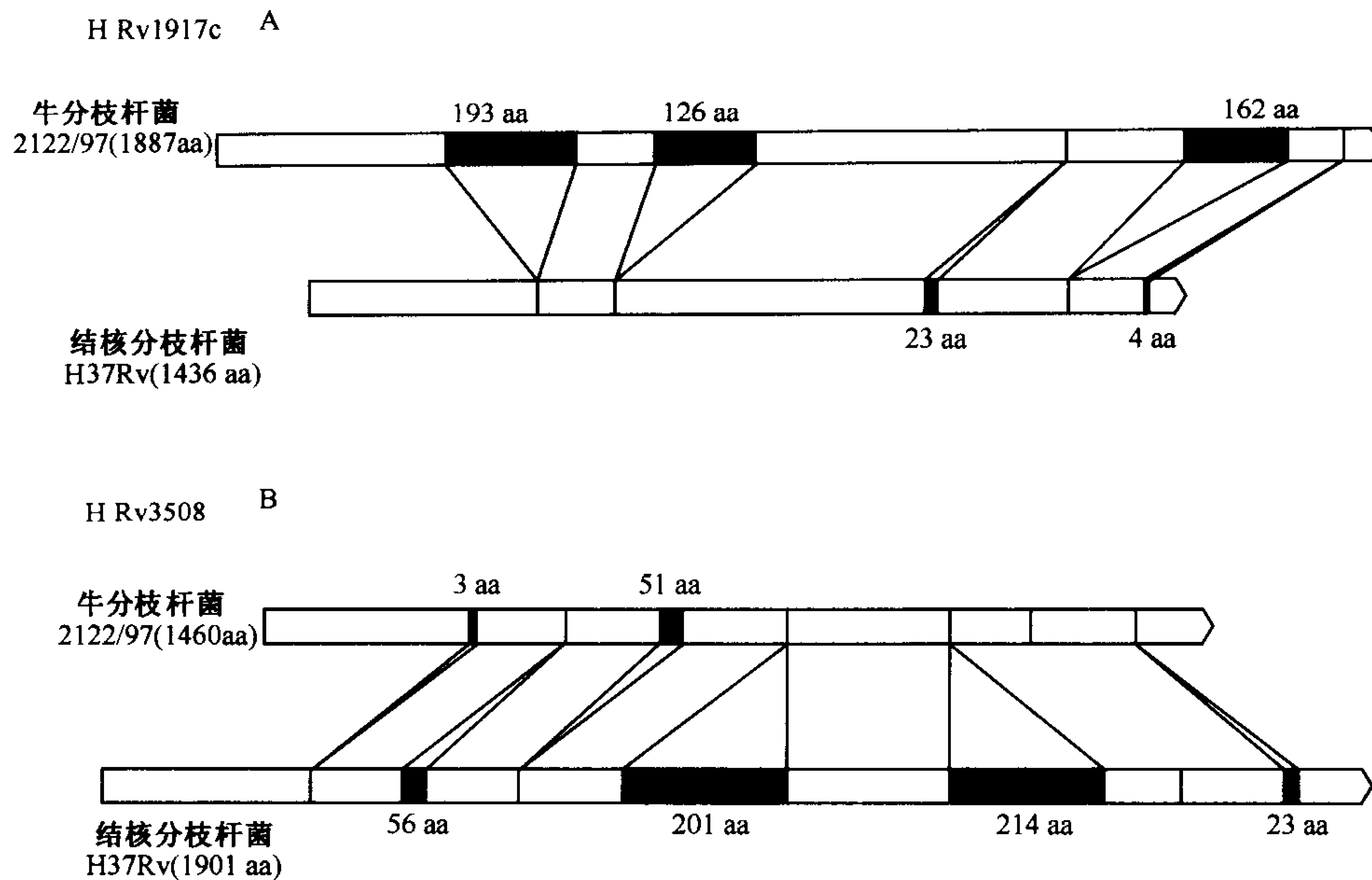


图 3 几种 PPE 基因的变化。结核分枝杆菌和牛分枝杆菌的直向同源基因及其插入位点和大小如图所示。

TbD1，它含有 *mmpS6* 基因及 *mmpL6* 的 5′端[42]（图 2）。Mmp 是一个跨膜蛋白家族，与参与细胞壁脂质输出的 RND 运输蛋白同源[43]，因此，牛分枝杆菌的 TbD1 位点也可能负责脂类物质的输出，如酚类糖脂，而此类脂质正是结核分枝杆菌所缺乏的。另外，与结核分枝杆菌相比，牛分枝杆菌的 *mmpL1* 和 *pks6* 基因的可读框发生了位移，Pks6 参与复合脂类物质的合成，而 MmpSL1 则参与这些脂类的输出，因此，这两个基因功能的丧失会对细胞壁产生影响。

代谢方面

在牛分枝杆菌中，关于无机硫代谢的突变属无害突变。*astA* 编码芳基硫酸酯酶，后者负责水解硫酯以释放无机硫，而牛分枝杆菌的 *astA* 不完整。类似情况还发生在无机硫的运输蛋白 GysTWA 上，牛分枝杆菌菌株 BCG 的 GysTWA 失活并不影响该菌在寄主体内生存能力[44]；而在麻风分枝杆菌（*M. leprae*）中，*cysTWA* 根本就是一个假基因（pseudogene）[3]。因此，这类细菌在寄主中的生存不需要摄取无机硫，而牛分枝杆菌在寄主体内的生存却要依赖有机硫（可能通过氨基酸转运酶来实现）。

treY 编码 MOT（maltooligosyltrehalose）合成酶，TreY 催化终端 α（1，4）葡萄糖键转变为 α-（1，1）键，随后 MOT 双糖进一步被海藻糖水解酶（由 *treZ* 编码）水解，释放出自由海藻糖。但牛分枝杆菌的 *treY* 基因缺失 808 bp（如图 2），尚不清楚 *treY* 缺失对表现型的影响，海藻糖合成的另外两条途径完整无缺，而且它们可能协同表达。此外，值得注意的是 *treY* 缺失紧挨着 TbD1（图 2），对各种菌株的 PCR 分析表明，尽管 *treY* 缺失存在于大多数菌株中，但牛分枝杆菌菌株 BCG 的 *treY* 却是完整的，表明这种缺失可作为种内进一步分类的遗传标记。

麻风分枝杆菌基因组序列

引言

在某些方面，麻风分枝杆菌（*M. leprae*）可认为是分枝杆菌属边缘的一个成员，如相对低的 G+C 含量（57%）、较小基因组（比其他的至少小 1.1 Mb）（表 1）。对难以研究的细菌，基因组方法无疑是重要的手段。麻风分枝杆菌不能在培养基中体外培养，在小鼠上的代时（generation time）极长，达 14 天之久。基因组序列分析揭示，有大量假基因或突变失活基因，可以清楚地解释这些特征[3]。

麻风分枝杆菌基因组序列

有一株仇蜍起源的印度麻风病杆菌，基因组大小为 3.27Mb，有 1605 个蛋白质编码基因，50 个 RNA 编码基因（表 1）。与结核分枝杆菌基因组相比，这是一个极端简并进化（reductive evolution）的例子，尽管有很多（至少 27%）假基因，但仍然只有将近一半（49.5%）的基因组编码功能基因。假基因表现为功能丧失，原因可能是由于一个或多个突变所致，包括（提前）产生终止密码、移码、缺失以及（少数情况下发生的）插入，其中麻风分枝杆菌 1116 个假基因的同源物在结核杆菌中有功能，实际数目肯定比这更多，因为有的基因已经突变得面目全非，不可识别，有的基因在结核杆菌中没有同源物，这些也许就是余下的 23.5% 基因组，表现为非编码序列或“空”序列。

这 1116 个可识别的假基因在基因组中随机分布，而 1605 个功能基因多以群簇形式存在，并且两侧多被非编码 DNA 包围，这种基因衰退与 G+C 含量的降低有关：预测功能基因的 G+C 含量为 60.1%，可识别假基因为 56.5%，而非编码区域经历了最严重的退化，其 G+C 含量只有 54.5%。对麻风分枝杆菌更广义对所有分枝杆菌，有效基因较高 G+C 含量是由于活跃基因对密码子的选择所造成，非编码区域的随机突变则导致产生与寄主更接近的中等 G+C 含量。DNA 链中胞嘧啶脱氨后变成胸腺嘧啶，可能是麻风分枝杆菌 G+C 含量在所有分枝杆菌中最低的原因。有人指出，经历简并进化的微生物基因组一般富含 A+T[45]。

麻风分枝杆菌的简并进化

简并进化在许多重要人类病原菌中都有研究记载，如严格细胞内寄生菌立克次体（*Rickettsia*）和衣原体（*Chlamydia* spp）[46]（参见本书第 17 章）。一般认为在特定小生境中，不需要的基因会逐渐失活，也就是说进入了进化的死胡同。*Buchnera* spp 与肠道细菌有些类似，它是蚜虫的共生体，其简并进化如此彻底，基因组由原来的约 4.5 Mb 缩减至 0.64 Mb[47]，几乎没有假基因，所以，缺失是基因组衰减的主要方式。在麻风分枝杆菌基因组序列问世前，斑疹伤寒病原普氏立克次体（*Rickettsia prowazekii*）是已报道基因组衰退最为严重的，其中仅有 76% 潜在编码序列有功能[48]。与麻风分枝杆菌相比，普氏立克次体（*R. prowazekii*）的基因丢失水平仅算一般，但这两个病原菌有一共同特点：与 *Buchnera* 相比，它们删除假基因的进度远远落后于基因失活的进度[47]。

有一个称为 Muller 的棘齿（Miller' s ratchet）的很雅致的假说解释简并进化，遗传物质或基因功能随机丢失往往导致适应性降低，部分原因是这些微生物缺乏有性生殖循环，而不能获取异源 DNA 修复其遗传损害，其直接结果是缺乏遗传多样性。显然由于栖息生境高度专一，与麻风分枝杆菌发生遗传物质交换的只能是其寄主：人类或上祖哺乳动物。有证据表明，微生物与寄主间曾经发生过基因水平转移，例证之一是 *proS*，它编码脯氨酰-tRNA 合成酶，麻风分枝杆菌的所有氨酰-tRNA 合成酶基因中，只有 *proS* 在结核分枝杆菌中没有严格的对应体，而麻风分枝杆菌 ProS 的结构域与真核生物颇为类似，如人类、酵母、果蝇和柏格多弗疏螺旋体（*Borrelia burgdorferi*）。有人认为，氨酰-tRNA 合成酶基因的水平转移经常发生，柏格多弗疏螺旋体也是从其寄主获得的 *proS*[49]。还有一个证据是，与结核分枝杆菌相邻基因比较，麻风分枝杆菌的 *proS* 被置换过，且为倒置，表明是近期获得的[3]。在结核分枝杆菌以及已测序的所有分枝杆菌中，酶的结构域都具有典型的原核生物特征，该菌可能一度也有原核生物的 *proS*，后来被寄主的 *proS* 所置换。

麻风分枝杆菌的重复 DNA

麻风分枝杆菌基因组中有 2% 重复序列，它可能介导发生基因组重组[50]，在其 26 个以上不同插入元件（IS）中，没有一个检测出功能。麻风分枝杆菌有 4 个家族的重复子，零散分布于基因组中，它们是 RLEP（37 个拷贝）、REPLEP（15 个拷贝）、LEP-REP（8 个拷贝）和 LEPRPT（5 个拷贝），都不具有可读框，却都在某种程度上展现出它们曾经是可转座片段，如 2383 bp 的 LEPREP 上有一个 54bp 回文倒置重复和末端 6bp 倒置重复（5′-CTAGTG）。

BLASTX 搜寻表明，这些重复子与假产碱假单胞菌（*Pseudomonas putida*）和根癌土壤杆菌（*Agrobacterium tumefaciens*）的转座酶十分相似，与真菌第二类内含子成熟酶相关蛋白也有相似性。尽管 RLEP 与已知转座子的序列没有相似性，但它多发生在基因的 3′端和几个假基因中，表明它曾具有转座能力。881bp 的 REPLEP 往往以一个 8bp 的倒置重复为边界，LEPRPT 长度为 1254 bp[50]。将结核分枝杆菌与麻风分枝杆菌的基因组序列比较，发现许多重复序列位于基因次序的非连续区，证据还表明，这些分散在基因组中多拷贝重复子之间的重组，导致了基因组共线性（synteny）丧失、基因倒置和基因组体积缩减[50]。

麻风分枝杆菌的简并进化与生物学

在结核分枝杆菌复合群和天蓝色链霉菌（*Streptomyces coelicolor*）中，广泛存在着基因的复制加倍，继而产生了很大的蛋白质家族及功能冗余[4,10,51]。麻风分枝杆菌的基因复制加倍水平约为 34%[52]，其中最大的功能性蛋白质群涉及到脂肪酸和聚酮的代谢和修饰、代谢物的转运、细胞囊的合成和基因调控。在结核分枝杆菌中的大致趋势也是如此，其 52% 基因是由于复制加倍或结构域拼接而产生[10]。如果把假基因也包括在内，麻风分枝杆菌显然曾有更多的基因冗余，简并进化导致许多基因的丧失和某些功能的选择性保留。

像结核杆菌一样，麻风分枝杆菌的最大蛋白质家族包括聚酮合成和脂肪酸代谢的相

关酶[52]，再次表明这一类蛋白质对这类生长缓慢分枝杆菌病原菌的重要性。但是，麻风分枝杆菌这套酶系远远没有结核杆菌的复杂和浩大，因为后者细胞囊含有更多样化的脂肪、糖脂和碳水化合物[17]。在结核分枝杆菌 H37Rv 中，第二和第三大家族的蛋白质是 167 个 PE 和 PPE 蛋白质，其中只有 12 个存在于麻风分枝杆菌中，尽管还有约 30 个假基因存在。在预测这些麻风分枝杆菌 PE 或 PPE 蛋白质中，没有一个含有多重 C 端重复序列，而多重 C 端重复序列可能参与抗原变异[4,24,25]，由于这类基因家族都有特别丰富的 G+C，所以它们的减少可能也是导致麻风分枝杆菌 G+C 含量较低和基因组体积较小的原因。

基因退化和基因组缩减使某些整个代谢途径从基因组中消失，特别是关于分解代谢及其调控与辅佐途径。当仔细检查那些在简并进化中丢失或失活的基因时，可以清楚地看到与达尔文进化论相符合的倾向，例如，基因组比较分析表明，结核分枝杆菌和天蓝色链霉菌（*S. coelicolor*）都有用硝酸盐作为最终电子受体进行无氧呼吸的能力。在这个反应中，由醌传递的电子经甲酸盐脱氢酶-N 转给甲酸盐，电子进一步为硝酸盐还原酶捕获，产生质子动力[4,51]。硝酸还原酶有 3 个亚基，分别为 *narGHI* 所编码；甲酸盐脱氢酶-N 也含有 3 个亚基；甲酸盐由丙酮酸经乙酰辅酶-A 产生[53]。两种酶都利用钼蛋白为辅助因子，后者的合成至少有 9 个 *moe/moa* 基因产物参与，钼由 *modABC* 编码的 ABC 转运蛋白从细胞外培养基中吸收[4]。麻风分枝杆菌含有硝酸还原酶和甲酸盐脱氢酶-N 相对应的假基因，也具有负责钼吸收转运并将其插入蝶呤的假基因。显然，随着利用甲酸-硝酸代谢途径能力的丧失，整个系统的基因不断积累突变并逐步退化，由于机体已经不需要它们了。麻风分枝杆菌的有氧呼吸途径也很有意思，整个呼吸链大大缩短，只剩下 NADH 氧化酶操纵子 *nuoA-N* 的 3′端，奇怪的是，天蓝色链霉菌虽然拥有完整的 *nuoA-N* 操纵子，同时还有另一份残缺拷贝[51]。

与分解代谢和能量产生途径相比，麻风分枝杆菌的合成代谢途径基本上保存完好，这表明之所以不能人工培养麻风分枝杆菌，不是缺乏某种特定氨基酸、维生素或其他因子，而是采用碳源和能源不同的缘故。麻风分枝杆菌可能只能利用非常有限的碳源，甚至是特化的碳源组合，从而在选择性的氧化还原条件下维持其碳代谢的平衡。

分枝杆菌的比较基因组学及进化线索

引言

很多年来，由于研究手段的缺乏，分枝杆菌的遗传学研究总是落后于链霉菌。但是，近 10 年却有很大的发展，例如，许多基于分枝杆菌噬菌体、质粒和转座子遗传系统的建立[54~58]，以及用于基因敲除高效载体的构建[59,60]。同时，由于基因组学的发展，关于分枝杆菌的遗传学知识也大大地丰富（表 1，表 2）。在 20 世纪 70～80 年代，分子遗传学研究主要局限于模式物种，而现在分枝杆菌已成为遗传研究最为透彻的群体之一，为基因组比较研究打下了基础。

结核分枝杆菌复合群的进化

如前所述，结核分枝杆菌复合群中的各菌系高度相关，它们大多导致哺乳动物的结

核病，只是寄主范围和对人的致病力方面有差异。结核分枝杆菌、*M. canettii* 和非洲分枝杆菌（*M. africanum*）是严格的人类病原菌，田鼠分枝杆菌（*M. microti*）主要是啮齿动物病原菌，而牛分枝杆菌则有很宽的寄主范围，它能感染包括人类在内的大多数哺乳动物。BCG 疫苗菌是最有名的无毒菌株，由 Calmette 和 Guerin 从牛分枝杆菌毒性株出发，经过 230 次系列传代后获得。BCG 疫苗虽然被广泛使用，但对其减毒机制仍然一无所知。这株卓越的活疫苗再也没有重获毒力，说明在其减毒过程中，发生了不可逆的基因组变化，如缺失，这种现象激发了许多基因组比较研究[31, 32, 61, 62]。综合这些研究结果，共有 18 个 RD（Region of Difference，差异区段）（大小从 0.3～12.7 kb，代表 120 个基因）存在于结核分枝杆菌菌株 H37Rv 中，而不存在于牛分枝杆菌菌株 BCG Pasteur 中，其中某些 RD 可能与疫苗和毒性株之间的表型差异有关。

对这些 RD 进行特征 PCR 分析，发现不只是 BCG 缺乏大多数上述 RD 片段，牛分枝杆菌其他菌株也缺乏它们；这反映了结核分枝杆菌和牛分枝杆菌在进化中的分歧，也反映了在 BCG 菌株减毒过程中基因组的修饰变化。例如，田鼠分枝杆菌、牛分枝杆菌和牛分枝杆菌菌株 BCG 都没有 RD7-RD10。如果仔细检查相邻序列，RD7-RD10 在这些菌株中由于基因缺失而导致丢失，这也排除了另一种 RD 产生是由于外源片段插入结核分枝杆菌的可能性[63]。

由于非洲分枝杆菌缺乏 RD9，同样的缺失也应该发生在与非洲分枝杆菌属同一进化支系的其他成员中，如田鼠分枝杆菌和牛分枝杆菌，而不应该在属于不同进化支系的结核分枝杆菌中。假定不同支系间罕有遗传物质的水平转移，基于同一复合群某些成员中的保守性遗传缺失，有人提出了结核分枝杆菌复合群各成员的进化图（图 4）[42]，从图中可以看出，与结核分枝杆菌相比，牛分枝杆菌和牛分枝杆菌菌株 BCG 作为单独支系的最后成员，已积累了所有的遗传缺失。

这个图与富含信息 SNP（单核苷酸多型性）的分布颇为一致，如 *pncA*（抗吡嗪酰胺基因）的 SNP 分布[35]，它还与如下发现十分吻合：牛分枝杆菌 AF2122/97 的基因组序列比结核分枝杆菌 H37Rv 小 66 kb 左右（表 1）。该图的重要性还表现在它对传统的假说提出了挑战，过去一直认为人类结核分枝杆菌来源于牛分枝杆菌，即牛结核杆菌在牲畜的驯养过程中跨越了物种障碍进入人类[64, 65]。根据 RD 和 SNP 分布，结核分枝杆菌各菌株要比牛分枝杆菌菌株更接近结核分枝杆菌复合群的共同祖先。非洲分枝杆菌（RD9）、*M. canettii*（RD7，RD8～RD10）和田鼠分枝杆菌（RD4、RD5、RD7-RD10、RD12、RD13）代表独立的一个支系，它们很可能由今天结核分枝杆菌的祖先进化而来再适应新寄主。尽管需要古病理学者进一步确证，但是，有好几个基因在 *M. canettii*、结核分枝杆菌和非洲分枝杆菌等人类病原菌中完整无缺，却在田鼠分枝杆菌和牛分枝杆菌动物病菌中或缺或残缺，这种发现表明，结核杆菌的共同祖先应该早已是人类病原菌了。

如图 2 所示，比较基因组学发现了 TbD1 区域，TbD1 不存在于结核分枝杆菌的所有分离株中，却存在于结核分枝杆菌复合群的其他成员中。基于这种发现，正在利用 PCR 方法来准确迅速地鉴定结核分枝杆菌菌系。

分枝杆菌进化概述

为了更广义地理解进化，比较已测序其他分枝杆菌（表 2）的基因组很有必要，有

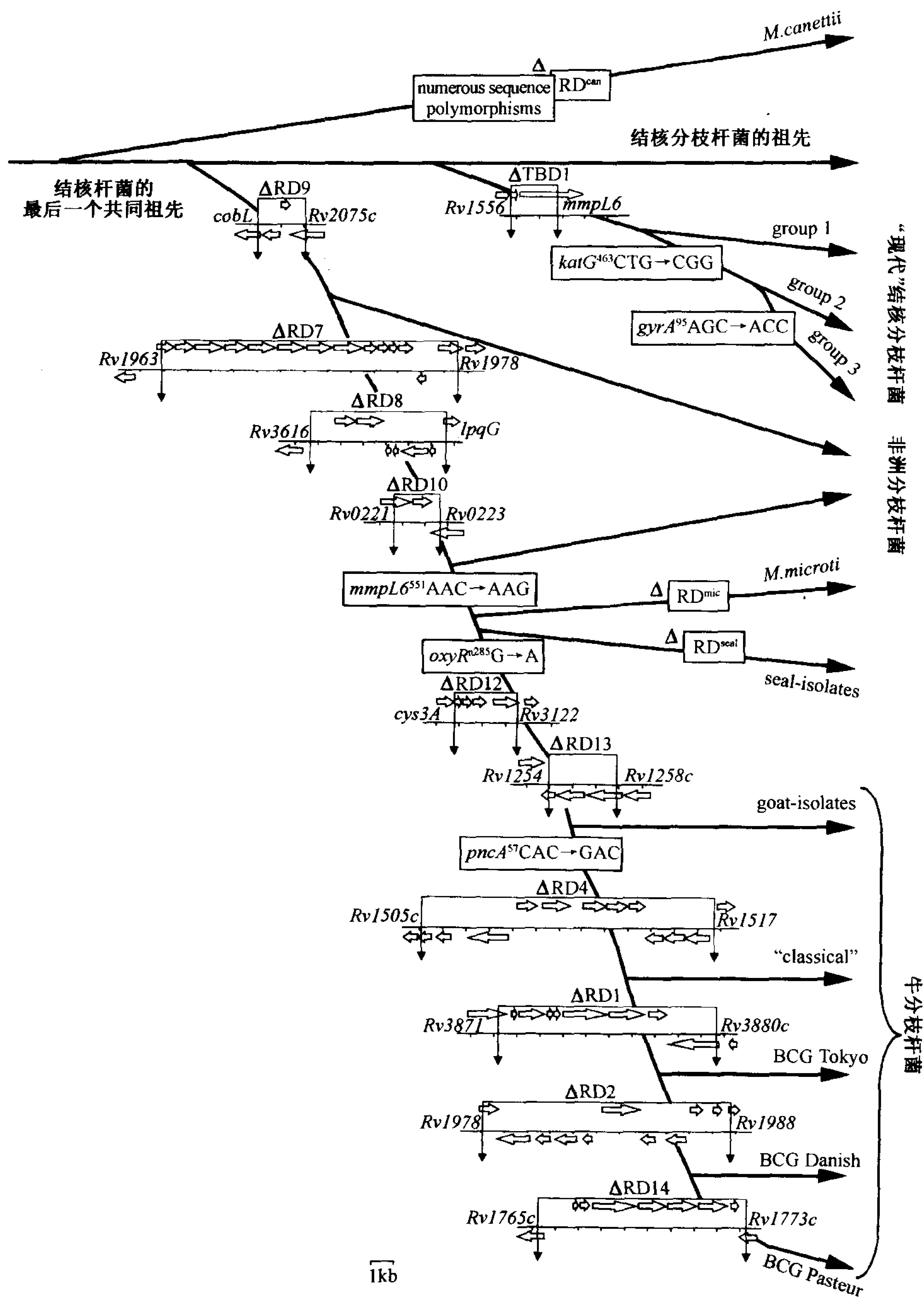


图 4　结核杆菌进化图。根据结核分枝杆菌复合群部分成员的保守缺失和特定 SNP；缺失区域由所示箭头界定；结核分枝杆菌 H37Rv 的可读框及其位置和转录方向由带箭头的方框所示。

助于理解为什么无害的土壤微生物会变为严格的细胞内寄生菌。根据种系进化，与结核分枝杆菌复合群最近的种群是环境菌株海分枝杆菌（*Mycobacterium marinum*）和溃疡分枝杆菌（*M. ulcerans*）。海分枝杆菌是变温动物的病原菌，几乎不感染人类，而溃疡

分枝杆菌则导致弱皮肤病（debilitating cutaneous disease），在西非一些地方有时流行。鸟型分枝杆菌（*M. avium*）复合群是另一类生长缓慢的分枝杆菌亚种，其中有牲畜病原菌和人类机会病原菌，鸟型分枝杆菌、副结核杆菌（*M. paratuberculosis*）和耻垢分枝杆菌（*M. smegmatis*）mc^2155 的基因组测序正在进行（表 2）。耻垢分枝杆菌（*M. smegmatis*）mc^2155 是生长迅速的一个种，最令人吃惊的是它的基因组大小为 7 Mb，与在进化上相距很远的放线菌天蓝色链霉菌[51]（在另一节中详细描述）相近，比那些生长缓慢的分枝杆菌基因组（~4.4 Mb）要大得多。

表 2　放线菌基因组相关网页

白喉杆菌(*Corynebacterium diphtheria*)
　http://www.sanger.ac.uk/Projects/C_diphtheriae/|
牛分枝杆菌(*Mycobcterium bovis*)
　http://www.sanger.ac.uk/Projects/M_bovis/
麻风分枝杆菌(*Mycobacterium leprae*)
　http://genolist.pasteur.fr/Leproma/
　http:/www.sanger.ac.uk/Projects/M_leprae/|
海分枝杆菌(*Mycobacterium marinum*)
　http://www.sanger.ac.uk/Projects/M_marinum/|
副结核杆菌(*Mycobacterium paratuberculosis*)
　http://www.cbc.umn.edu/Research Projects/AGAC/Mptb/Mptbhome.html
耻垢分枝杆菌(*Mycobacterium smegmatis*)，鸟型分枝杆菌（*Mycobacterium avium*）
　http://www.tigr.org/tdb/mdb/mdbinprogress.html
　http://tigrblast.tigr.org/ufmg/
结核分枝杆菌(*Mycobacterium tuberculosis*)
　H37Rv 菌系
　http://genolist.pasteur.fi/TubercuList/
　http://www.sanger.ac.uk/Projects/M_tuberculosis/
　CDC1551 菌系
　http://www.tigr.org/tigr-scripts/CMR2/
GenomePage3.spl? database=gmt
　210 菌系(北京型)
　http://tigrblast.tigr.org/ufmg/
溃疡分枝杆菌(*Mycobacterium ulcerans*)
　http://www.pasteur.fr/recherche/unites/Lgmb/
天蓝色链霉菌(*Streptomyces coelicolor*)
　http:www.sanger.ac.uk/Projects/S_coelicolor/
　http:jiio16.jic.bbsrc.ac.uk/S.coelicolor/

耻垢分枝杆菌的大基因组，加上基于 16S rRNA 的分类树表明，生长缓慢的分枝杆菌是该属中最近期进化的一类[66]，只有为数不多的几个基因，如 mgtC[67]，仅在缓慢生长而不在快速生长的分枝杆菌中。因此，缓慢生长的分枝杆菌变成病原菌的重要原因，似乎是遗传材料的丢失而不是获得。如前所述，这个趋势还在继续，一个极端的例

子就是麻风分枝杆菌基因组[3]。

比较基因组学在功能方面的发现

比较基因组学除了完全彻底地修订了结核杆菌进化史外[42]，还提供了新知识，以增进认识 BCG 免疫菌株最初减毒过程及其他与毒力相关的问题。最令人感兴趣的区段是 RD1 位点（图 4），RD1 存在于已经测试所有结核分枝杆菌和牛分枝杆菌的毒性株中[42]，唯独在所有 BCG 菌株中没有[31, 61]。RD1 编码高效 T 细胞抗原 CFP-10 和 ESAT-6，功能基因组学认为 BCG 失毒可能是 RD1 位点的丢失。

比较基因组学不仅揭示了遗传缺失，在菌株 BCG Pasteur 中，还发现了两个大串联重复区（DU1 和 DU2，分别为 29 和 36 kb）[68]，其位点和大小因不同 BCG 株系（sub-strain）而异，表明它们是彼此独立发生的。DU1 似乎仅存在菌株 BCG Pasteur 中，而 DU2 则存在所有供测的 BCG 株系中（作者未发表资料）。有趣的是，DU1 含有 *oriC* 位点，表明菌株 BCG Pasteur 的 *oriC* 及有关复制的基因是双倍体[68]；而 DU2 的串联重复导致 30 个基因为双倍体，其中包括与热激反应有关的 sigma 因子—*sigH*[69]。在实验条件下，当细菌面临各种选择压力时，基因复制加倍是进化的普通反应，由此推论在自然环境下也是如此[70]。因为复制加倍增加了基因量，能造成复制体两端发生基因融合而产生新功能，并能提供更丰富的遗传多样性资源。在结核杆菌和麻风杆菌中，这种基因复制加倍都有发生，耻垢分枝杆菌的一个 250kb 片段也像是加倍过[71]。这种复制加倍在放线菌中是否广泛存在，值得进一步研究。

天蓝色链霉菌基因组序列

链霉菌引言

链霉菌对土壤生境的适应非常成功，它们在土壤中几乎是无处不在。在营养、生理和生物学方面，土壤是非常复杂多变的环境，链霉菌栖息在这样的环境中，靠分枝状菌丝集结成菌体生长，依据菌丝状营养体的特点，链霉菌曾一度划为真菌。由于某些条件的改变，很可能是营养耗竭导致生长缓慢，菌丝体分化产生气生菌丝，其上是卷曲状孢子链（即产孢单元），这种现象常见于链霉菌生长的土壤颗粒表面。天蓝色链霉菌在链霉菌中研究得最详尽，因此是基因组测序的理想对象[72]。

大染色体，众多基因

天蓝色链霉菌基因组的显著特征是它的大小和结构，像大多数链霉菌一样，它有一个大线型染色体，长8 667 507bp，在已测序的细菌中，它的基因组最大[51]。具有典型细菌基因组编码密度为 88.9%，含有 7825 个基因，仅有 55 个假基因，见表 1。其基因组是结核分枝杆菌的两倍，比低等真核生物裂殖酵母（*Schizosaccharomyces pombe*）多 3000 个基因[73]，是预测人类基因数的四分之一[74,75]。这种基因数目过剩的现象有两个问题，为什么天蓝色链霉菌有这么多基因？这么多基因编码些什么？考虑到天蓝色链霉菌的生活史及其复杂生境，不难回答这两个问题。

撇开行使必需功能的管家（housekeeping）蛋白，天蓝色链霉菌将其余蛋白质组的

大部分用于对付细胞壁外面的事务。由于动物、植物、昆虫、真菌、细菌等各种生物腐烂产生的生物高聚物残留在土壤中，它富集了各种营养来源，为了利用各种营养，天蓝色链霉菌分泌819种蛋白质（占基因组10.5%）到细胞外，大部分是水解酶，包括蛋白酶、几丁质酶、纤维素酶、淀粉酶和果胶酶，这些水解反应的产物与金属、其他离子、氨基酸和多肽一起被运输到细胞质，因此，天蓝色链霉菌蛋白组另一大功能就是搬进送出，这类蛋白质共614个，占基因组7.8%。

天蓝色链霉菌最称著的输出产物是抗生素，这些化合物由次生代谢途径合成，能参与各种各样次生代谢物质合成的基因簇有22个[51]，其中仅4个基因簇是在基因组测序前已知的，其中3个编码抗生素合成的酶类[76]，1个负责孢子色素的产生[77]，另外18个基因簇的发现使参与次生代谢的基因数目达到220个左右（有些基因簇的边界有待实验确证）。显然，次生代谢对天蓝色链霉菌不是偶然，如此规模基因组的投资，必将产生相当的竞争优势。产生钙依赖性抗生素的基因簇有41个[78]，该基因簇的投资回报可能产生对其性竞争细菌有抑制作用的化合物。新发现基因簇的预测功能包括：抗干旱、适应低温和利用环境中的铁。

与结核分枝杆菌类似，天蓝色链霉菌有一整套细胞色素P450基因，由基因组序列预测，有18种这样的酶，并且都经过在大肠杆菌中表达验证[79]，这为天蓝色链霉菌提供了另一种潜在的适应性，去利用土壤中的各种有机物质。

为了管理庞大的基因队伍，天蓝色链霉菌基因组拥有空前规模的调控蛋白，占基因组12.3%，这与细菌的调控基因数目与其基因组大小成正比例的一般原则相符[80]。该菌还具有很多胞质外sigma因子[81]和大量（53个）双组分感应/调控蛋白对（two-component sensor/regulator pair），表明它要探测胞外环境，并作出相应的反应，这对天蓝色链霉菌有非常重要的意义。除了已知的调控蛋白外，有一类DNA结合蛋白（25个）是天蓝色链霉菌所独有的，可能代表新调控因子。

双相染色体结构

可以将天蓝色链霉菌的线状染色体分为三个区段：核心区和两个侧臂。核心区起源所有放线菌的共同祖先，侧臂的来源则不同，图5是天蓝色链霉菌与结核分枝杆菌的基因组比较，图中每个点代表两物种相应的最佳匹配蛋白质以及它们在染色体上的相对位置；图中不完整X形[82]反映了两种染色体之间的共线性和基因序列与先后顺序的保守性。

这种共线区散布于结核分枝杆菌整个染色体，却只集中于天蓝色链霉菌的核心区，位于共线区的基因主要涉及细胞基本功能，如细胞分裂、核苷酸与氨基酸的合成、中心代谢及核糖体的合成。这一现象与天蓝色链霉菌染色体的双相结构有关，几乎所有的必要基因都位于核心区，其他非必需基因通常位于侧臂上，如产生次生代谢物或某种特定胞外水解酶的基因。在实验室条件下，链霉菌可以忍受染色体末端1Mb或更多DNA缺失[83]，也许在有利条件下，核心区本身对链霉菌的生长繁殖已经足够，侧臂作为备用基因库只在特定环境条件下发挥作用。

天蓝色链霉菌的双相染色体结构能提供一些线索，帮助了解链霉菌线状大染色体是如何进化的。假定放线菌共同祖先的染色体是环状[84]，后来在某一时间，它获得了以

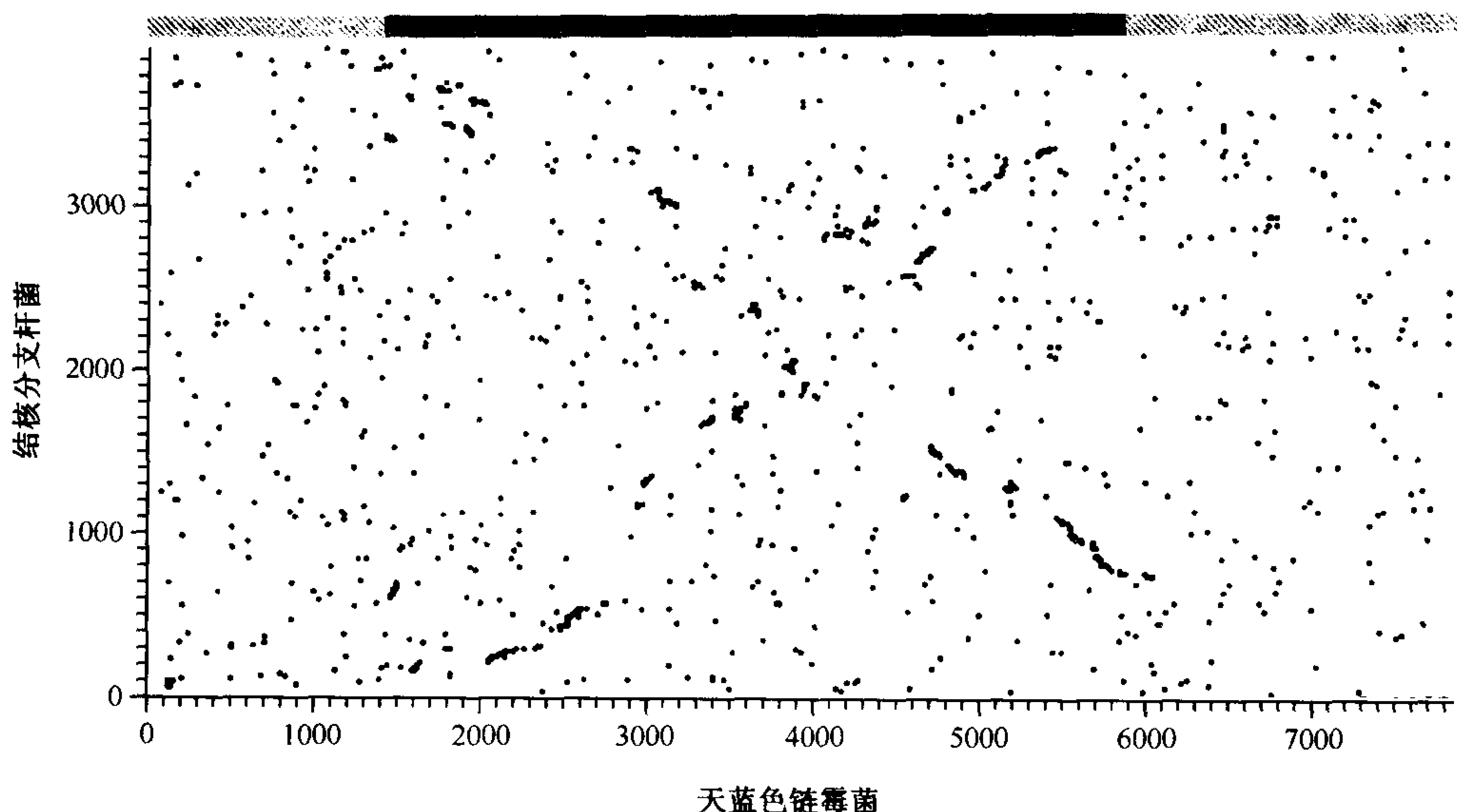


图 5　结核分枝杆菌（*M. tuberculosis*）和天蓝色链霉菌（*S. coelicolor*）的染色体结构比较。纵横坐标代表（编码蛋白质的）基因在染色体上的相对位置。DnaA 位于每个基因组中心。小圆点代表蛋白质两两比较的最佳匹配［经由 FASTA（fast-all）比较］。图上方的实线及阴影横条分别代表天蓝色链霉菌（*S. coelicolor*）染色体的核心区（SCO1440-5869）和侧臂。

线状方式生存的能力。也许是通过线性质粒整合或与线性质粒重组，该质粒可以产生稳定的线性 DNA 末端或链霉菌中常见的细菌“端粒”[85]。天蓝色链霉菌的亚端粒区含有较高比例的转座酶和假基因，表明该区对遗传插入有增强的容忍性。也许新的线性祖先具有相似的容忍性，进而为末端特异性 DNA 的累积提供了一条途径。

假如获得的基因产生了竞争优势，就会选择性地保留下来，导致基因组体积的增加。天蓝色链霉菌的侧臂不断扩增，以至几乎占半个基因组。这样大的额外基因库，使天蓝色链霉菌的染色体侧臂含有看似浪费的一个部件，一个例子是在两个相对侧臂上，有两个独立操纵子，都具有编码合成气囊的潜力[51]；另一个例子是零散分布的 13 个保守子（conservon），它们是 4 基因操纵子（功能未知）的各种变异拷贝，这类功能未知的操纵子只在结核分枝杆菌以单一拷贝出现过一次（Rv3362c～Rv3365c）。黏细菌（*Myxobacteria*）有更大的环状基因组[84]，表明线性并不是大染色体存在的先决条件，但是无能如何，链霉菌线性染色体为基因组扩增提供了有效途径。

放线菌比较基因组学

如前所述，比较基因组学正在揭示放线菌特有的保守基因和操纵子，它们将会在未来几年内广泛深入研究。尽管还是处于初步阶段，有些主题显而易见，在分枝杆菌中属 ESAT-6 簇类的蛋白质，由于具有 T 细胞抗原性和作为亚单位疫苗或诊断试剂的潜在性，而激起了很大的研究兴趣。在结核分枝杆菌 H37Rv 中，有 11 对 ESAT-6 基因的遗传排列显示，它们的蛋白质之间可能存在相互作用[39]。

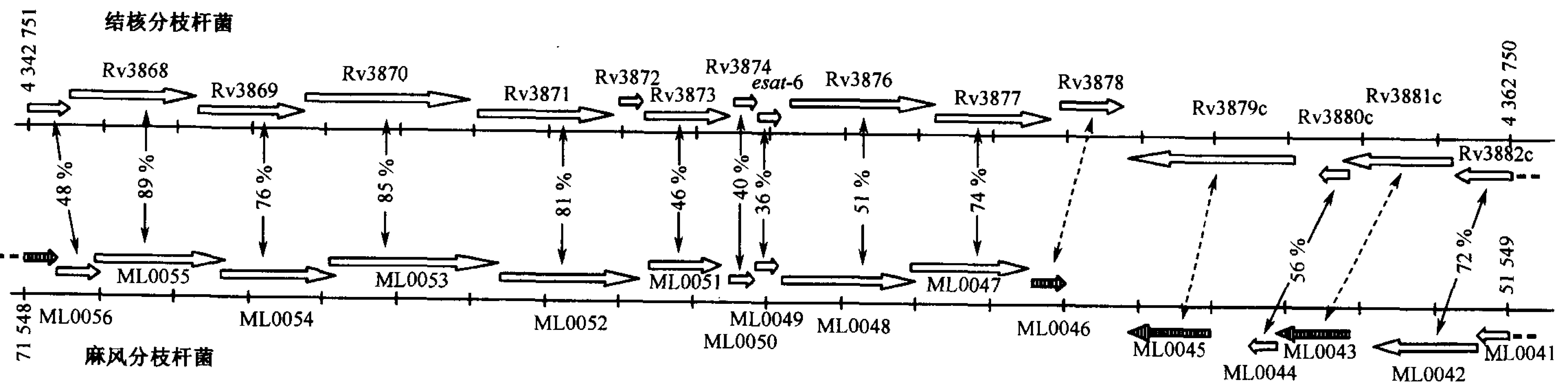


图6 结核分支杆菌(*M. tuberculosis*)和麻风分枝杆菌(*M. leprae*) ESAT-6位点的保守结构。空箭头代表开放阅读框架及其转录方向；箭头之间的百分数表示两个直向同源物之间在氨基酸水平上的相似程度。图中的数字表明开放阅读框架在基因组中的位置。注意，相对于结核分枝杆菌(*M. tuberculosis*)，麻风分枝杆菌(*M. leprae*) ESAT-6操纵子位于互补链上，在*oriC*的另一边。假基因用阴影箭头表示。

编码原型 ESAT-6 的基因[38]位于 RD1 位点，靠近结核杆菌的复制起点[61]。尽管在麻风分枝杆菌中基因褪变非常严重，但是，位于结核分枝杆菌 RD1 基因在麻风分枝杆菌中还是保守的[3]（图 6），可见它们在分枝杆菌生物学中是何等重要。还有，ESAT-6 基因不仅局限于致病分枝杆菌。虽然耻垢分枝杆菌和海分支枝菌的基因组测序正在进行中，但生物学搜寻表明它们的基因组中存在着 ESAT-6 的直系同源物（ortholog），在更远亲的白喉棒状杆菌（*C. diphtheriae*）和天蓝色链霉菌中，都发现有 ESAT-6 成员，尽管同源性比较低[86]，却说明了 ESAT-6 的重要性。

靠近复制原点 *oriC* 有一群 5 个保守基因（图 7），它们编码的蛋白质在分枝杆菌、棒杆菌和链霉菌中参与信号传导和细胞分裂[3, 4, 87, 88]。在所有情况下，它们的转录与复制方向相反，在天蓝色链霉菌中，这个假定操纵子只有一个基因，编码丝氨酸-苏氨酸含量偏低的蛋白质激酶[51]。

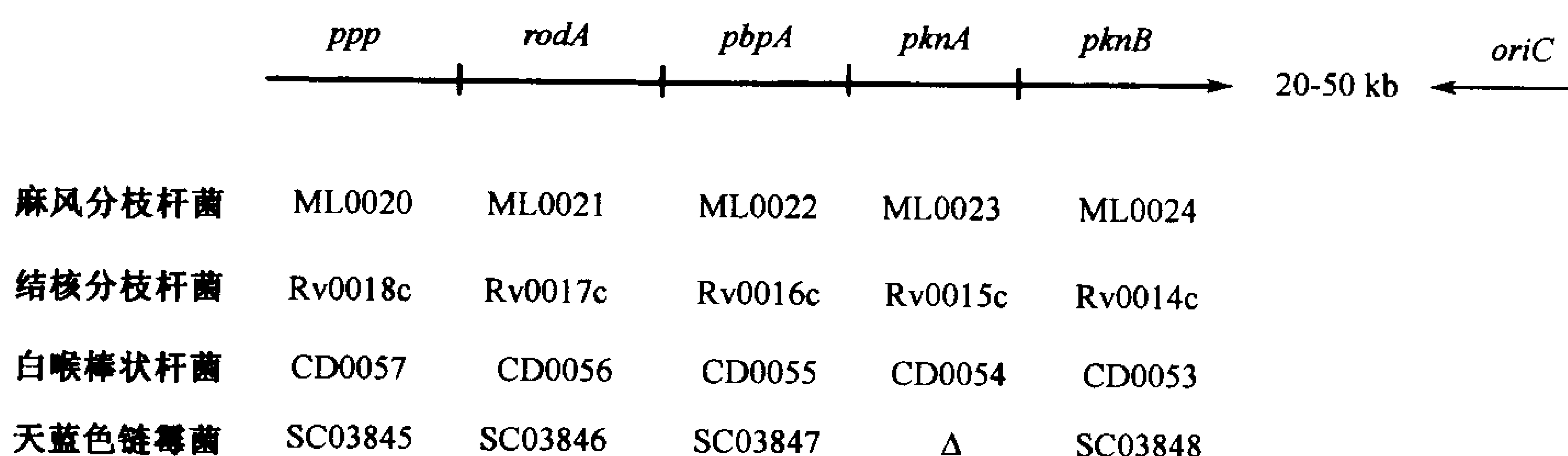


图 7　放线菌 *ppp-pknB* 基因簇的保守结构。该基因簇含 5 个基因，天蓝色链霉菌（*S. coelicolor*）是例外，它缺少 *pknA*；在所有 4 个种中，该基因簇都位于 *oriC* 附近，只是转录方向跟复制方向相反；基因符号如下：*ppp*，磷蛋白磷酸酯酶（phosphoprotein phosphatase）；*rodA*，细胞分裂蛋白；*pbpA*，肽聚糖生物合成蛋白；*pknA*，*pknB*，丝氨酸-苏氨酸-蛋白激酶。

放线菌另一个不寻常特征是它的 DNA 错配修复系统。与革兰氏阴性菌不同，放线菌好像缺乏正常行使修复功能的 *mutL-mutS* 系统[89]，但是却含有多个 *mutT* 直系同源物，毫无疑问，进一步比较分析会揭示更多值得仔细研究其功能的基因组区域。

结语

这一章我们叙述了放线菌基因组学方面的惊人进展，强调这些新知识新信息如何帮助认识放线菌这一博大群体的生物学并如何促进它们功能的开发。随着更多基因组序列问世，我们期待对种系发生（phylogeny）有更深的认识，对行为和发育过程有丰富的洞察，对区别放线菌与其他原核生物有更清楚的明确指标。

致谢

我们感谢放线菌基因组学同事们的贡献，并感谢以下单位或组织在财力上的帮助：Institut Pasteur, the Association Francaise Raoul Follereau, ILEP, the New York Community Trust, the Wellcome Trust, the European Union (QLRT-2000-02018), BBSRC 和 DEFRA。

（王清锋　译）

参考文献

1. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990; 87:4576–4579.
2. Stackebrandt E, Rainey FA, Ward-Rainey NL. Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. Int J Syst Bacteriol 1997; 47:479–491.
3. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001; 409:1007–1011.
4. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998; 393:537–544.
5. Sreevatsan S, Pan X, Stockbauer KE, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci USA 1997; 94:9869–9874.
6. Kapur V, Whittam TS, Musser J. Is *Mycobacterium tuberculosis* 15,000 years old? J Infect Dis 1994; 170:1348–1349.
7. Camus J-C, Pryor MJ, Médigue C, Cole ST. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiology 2002; 148(Pt 10):2967–2973.
8. Fleischmann RD, Alland D, Eisen JA, et al. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 2002; 184:5479–5490.
9. Valway SE, Sanchez MP, Shinnick TF, et al. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. N Engl J Med 1998; 338:633–639.
10. Tekaia F, Gordon SV, Garnier T, Brosch R, Barrell BG, Cole ST. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuber Lung Dis 1999; 79:329–342.
11. Gordon SV, Heym B, Parkhill J, Barrell B, Cole ST. New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. Microbiology 1999; 145:881–892.
12. Lee TY, Lee TJ, Belisle JT, Brennan PJ, Kim SK. A novel repeat sequence specific to *Mycobacterium tuberculosis* complex and its implications. Tuber Lung Dis 1997; 78:13–19.
13. van Embden JDA, Cave WM, Crawford JT, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993; 31:406–409.
14. Fang Z, Doig C, Kenna DT, et al. IS*6110*-mediated deletions of wild-type chromosomes of *Mycobacterium tuberculosis*. J Bacteriol 1999; 181:1014–1020.
15. Brosch R, Philipp W, Stavropolous E, Colston MJ, Cole ST, Gordon SV. Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra. Infect Immun 1999; 67:5768–5774.
16. Ho TBL, Robertson BD, Taylor GM, Shaw RJ, Young DB. Comparison of *Mycobacterium tuberculosis* genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. Comparative Functional Genomics (Yeast) 2001; 17:272–282.
17. Daffe M, Draper P. The envelope layers of mycobacteria with reference to their pathogenicity. Adv Microb Physiol 1998; 39:131–203.
18. Peterson JA, Graham SE. A close family resemblance: the importance of structure in understanding cytochromes P450. Structure 1998; 6:1079–1085.
19. Aoyama Y, Horiuchi T, Gotoh O, Noshiro M, Yoshida Y. CYP51-like gene of *Mycobacterium tuberculosis* actually encodes a P450 similar to eukaryotic CYP51. J Biochem 1998; 124: 694–696.

20. Podust LM, Poulos TL, Waterman MR. Crystal structure of cytochrome P450 14alpha-sterol demethylase (CYP51) from *Mycobacterium tuberculosis* in complex with azole inhibitors. Proc Natl Acad Sci USA 2001; 98:3068–3073.
21. Bellamine A, Mangla AT, Nes WD, Waterman MR. Characterization and catalytic properties of the sterol 14alpha-demethylase from *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 1999; 96:8937–8942.
22. Weber I, Fritz C, Ruttkowski S, Kreft A, Bange FC. Anaerobic nitrate reductase (narGHJI) activity of *Mycobacterium bovis* BCG in vitro and its contribution to virulence in immunodeficient mice. Mol Microbiol 2000; 35:1017–1025.
23. Cole ST. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. FEBS Lett 1999; 452:7–10.
24. Banu S, Honoré N, Saint-Joanis B, Philpott D, Prévost M-C, Cole ST. Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? Mol Microbiol 2002; 44: 9–19.
25. Brennan MJ, Delogu G. The PE multigene family: a molecular mantra for mycobacteria. Trends Microbiol 2002; 10:246–249.
26. Brennan MJ, Delogu G, Chen Y, et al. Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. Infect Immun 2001; 69: 7326–7333.
27. Delogu G, Brennan MJ. Comparative immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. Infect Immun 2001; 69:5606–5611.
28. Espitia C, Laclette JP, Mondragon-Palomino M, et al. The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? Microbiology 1999; 145:3487–3495.
29. Camacho LR, Ensergueix D, Perez E, Gicquel B, Guilhot C. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. Mol Microbiol 1999; 34:257–267.
30. Ramakrishnan L, Federspiel NA, Falkow S. Granuloma-specific expression of mycobacterium virulence proteins from the glycine-rich PE-PGRS family. Science 2000; 288:1436–1439.
31. Behr MA, Wilson MA, Gill WP, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarrays. Science 1999; 284:1520–1523.
32. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol 1999; 32:643–656.
33. Titball RW. Bacterial phospholipases. Symp Ser Soc Appl Microbiol 1998; 27:127S–137S.
34. Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley LW. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. Science 1993; 261:1454–1457.
35. Scorpio A, Zhang Y. Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. Nat Med 1996; 2:662–667.
36. Collins DM, Kwakami RP, de Lisle GW, Pascopella L, Bloom BR, Jacobs JWR. Mutation of the principal s factor causes loss of virulence in a strain of the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci USA 1995; 92:8036–8040.
37. Wiker HG, Nagai S, Hewinson RG, Russell WP, Harboe M. Heterogenous expression of the related MPB70 and MPB83 proteins distinguish various substrains of *Mycobacterium bovis* BCG and *Mycobacterium tuberculosis* H37Rv. Scand J Immunol 1996; 43:374–380.
38. Sorensen AL, Nagai S, Houen G, Andersen P, Andersen Å. Purification and characterization of a low molecular mass T-cell antigen secreted by *Mycobacterium tuberculosis*. Infect Immun

1995; 63:1710–1717.
39. Renshaw PS, Panagiotidou P, Whelan A, et al. Conclusive evidence that the major T-cell antigens of the *M. tuberculosis* complex ESAT-6 and CFP-10 form a tight, 1:1 complex and characterisation of the structural properties of ESAT-6, CFP-10 and the ESAT-6-CFP-10 complex: implications for pathogenesis and virulence. J Biol Chem 2002; 8:8.
40. Li J, Ochman H, Groisman EA, et al. Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica*. Proc Natl Acad Sci USA 1995; 92:7252–7256.
41. Sokurenko EV, Chesnokova V, Dykhuizen DE, et al. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. Proc Natl Acad Sci USA 1998; 95:8922–8926.
42. Brosch R, Gordon SV, Marmiesse M, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci USA 2002; 99:3684–3689.
43. Tseng T-T, Gratwick KS, Kollmann J, et al. The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and develoment proteins. J Mol Microbiol Biotechnol 1999; 1:107–125.
44. Wooff E, Michell SL, Gordon SV, et al. Functional genomics reveals the sole sulphate transporter of the *Mycobacterium tuberculosis* complex and its relevance to the acquisition of sulphur in vivo. Mol Microbiol 2002; 43:653–663.
45. Tamas I, Klasson LM, Sandstrom JP, Andersson SG. Mutualists and parasites: how to paint yourself into a (metabolic) corner. FEBS Lett 2001; 498:135–139.
46. Andersson JO, Andersson SGE. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev 1999; 9:664–671.
47. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. Nature 2000; 407:81–86.
48. Anderssen SGE, Zomorodipour A, Andersson JO, et al. The complete genome sequence of the obligate intracellular parasite *Rickettsia prowazekii*. Nature 1998; 396:133–140.
49. Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of amino-acyl–tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 1999; 9:689–710.
50. Cole ST, Supply P, Honoré N. Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity. Lepr Rev 2001; 72:449–461.
51. Bentley SD, Chater KF, Cerdeno-Tarraga AM, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 2002; 417:141–147.
52. Eiglmeier K, Parkhill J, Honoré N, et al. The decaying genome of *Mycobacterium leprae*. Lepr Rev 2001; 72:387–398.
53. Jormakka M, Tornroth S, Byrne B, Iwata S. Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. Science 2002; 295:1863–1868.
54. Jacobs WR Jr, Kalpana GV, Cirillo JD, et al. Genetic systems for mycobacteria. Methods Enzymol 1991; 204:537–555.
55. Snapper SB, Lugosi L, Jekkel A, et al. Lysogeny and transformation in mycobacteria: stable expression of foreign genes. Proc Natl Acad Sci USA 1988; 85:6987–6991.
56. Snapper SB, Melton RE, Mustafa S, Kieser T, Jacobs WR. Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis*. Mol Microbiol 1990; 4: 1911–1919.
57. Bardarov S, Kriakov J, Carriere C, et al. Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 1997; 94: 10,961–10,966.
58. Pelicic V, Jackson M, Reyrat JM, Jacobs WR Jr, Gicquel B, Guilhot C. Efficient allelic exchange

and transposon mutagenesis in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 1997; 94: 10,955–10,960.

59. Parish T, Stoker NG. Use of a flexible cassette method to generate a double unmarked *Mycobacterium tuberculosis tlyA plcABC* mutant by gene replacement. Microbiology 2000; 146: 1969–1975.
60. Hinds J, Mahenthiralingam E, Kempsell KE, et al. Enhanced gene replacement in mycobacteria. Microbiology 1999; 145:519–527.
61. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. J Bacteriol 1996; 178: 1274–1282.
62. Salamon H, Kato-Maeda M, Small PM, Drenkow J, Gingeras TR. Detection of deleted genomic DNA using a semiautomated computational analysis of GeneChip data. Genome Res 2000; 10: 2044–2054.
63. Brosch R, Pym AS, Gordon SV, Cole ST. The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends Microbiol 2001; 9:452–458.
64. Bates JH, Stead WW. The history of tuberculosis as a global epidemic. Med Clin North Am 1993; 77:1205–1217.
65. Stead WW, Eisenach KD, Cave MD, et al. When did *Mycobacterium tuberculosis* infection first occur in the New World? An important question with public health implications. Am J Respir Crit Care Med 1995; 151:1267–1268.
66. Pitulle C, Dorsch M, Kazda J, Wolters J, Stackebrandt E. Phylogeny of rapidly growing members of the genus *Mycobacterium*. Int J Syst Bacteriol 1992; 42:337–343.
67. Buchmeier N, Blanc-Potard A, Ehrt S, Piddington D, Riley L, Groisman EA. A parallel intraphagosomal survival strategy shared by *Mycobacterium tuberculosis* and *Salmonella enterica*. Mol Microbiol 2000; 35:1375–1382.
68. Brosch R, Gordon SV, Buchrieser C, Pym A, Garnier T, Cole ST. Comparative genomics uncovers tandem chromosomal duplications in some strains of *Mycobacterium bovis* BCG: implications for vaccination. Comparative Functional Genomics (Yeast) 2000; 17:111–123.
69. Fernandes ND, Wu QL, Kong D, Puyang X, Garg S, Husson RN. A mycobacterial extracytoplasmic sigma factor involved in survival following heat shock and oxidative stress. J Bacteriol 1999; 181:4266–4274.
70. Lupski JR, Roth JR, Weinstock GM. Chromosomal duplications in bacteria, fruit flies, and humans. Am J Hum Genet 1996; 58:21–27.
71. Galamba A, Soetaert K, Wang XM, De Bruyn J, Jacobs P, Content J. Disruption of *adhC* reveals a large duplication in the *Mycobacterium smegmatis* mc(2)155 genome. Microbiology 2001; 147: 3281–3294.
72. Redenbach M, Kieser HM, Denapaite D, et al. A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. Mol Microbiol 1996; 21:77–96.
73. Wood V, Gwilliam R, Rajandream MA, et al. The genome sequence of *Schizosaccharomyces pombe*. Nature 2002; 415:871–880.
74. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291: 1304–1351.
75. Consortium IHGS. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860–921.
76. Hopwood DA, Chater KE, Bibb MJ. Genetics of antibiotic production in *Streptomyces coelicolor* A3(2), a model streptomycete. In: Vining LC, Stuttard C. (eds). Genetics and Biochemistry of Antibiotic Production. Newton, MA: Butterworth-Heinemann, 1995, pp. 65–102.

77. Davis NK, Chater KF. Spore colour in *Streptomyces coelicolor* A3(2) involves the developmentally regulated synthesis of a compound biosynthetically regulated to polyketide antibiotics. Mol Microbiol 1990; 4:1679–1691.
78. Chong PP, Podmore SM, Kieser HM, et al. Physical identification of a chromosomal locus encoding biosynthetic genes for the lipopeptide calcium-dependent antibiotic (CDA) of *Streptomyces coelicolor* A3(2). Microbiology 1998; 144:193–199.
79. Lamb DC, Skaug T, Song HL, et al. The cytochrome P450 complement (CYPome) of *Streptomyces coelicolor* A3(2). J Biol Chem 2002; 9:9.
80. Stover CK, Pham XQ, Erwin AL, et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 2000; 406:959–964.
81. Lonetto MA, Brown KL, Rudd KE, Buttner MJ. Analysis of the *Streptomyces coelicolor sigE* gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions. Proc Natl Acad Sci USA 1994; 91:7573–7577.
82. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 2000; 1, Genome Biology (RESEARCH) 0011.
83. Volff JN, Altenbuchner, J. Genetic instability of the *Streptomyces* chromosome. Mol Microbiol 1998; 27:239–246.
84. Casjens S. The diverse and dynamic structure of bacterial genomes. Annu Rev Genet 1998; 32: 339–377.
85. Chen CW, Huang CH, Lee HH, Tsai HH, Kirby R. Once the circle has been broken: dynamics and evolution of Streptomyces chromosomes. Trends Genet 2002; 18:522–529.
86. Gey Van Pittius NC, Gamieldien J, Hide W, Brown GD, Siezen RJ, Beyers AD. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. Genome Biol 2001; 2, RESEARCH0044.
87. Fsihi H, De Rossi E, Salazar L, et al. Gene arrangement and organisation in a ~76 kilobase fragment encompassing the *oriC* region of the chromosome of *Mycobacterium leprae*. Microbiology 1996; 142:3147–3161.
88. Av-Gay Y, Davies J. Components of eukaryotic-like protein signaling pathways in *Mycobacterium tuberculosis*. Microb Comp Genomics 1997; 2:63–73.
89. Mizrahi V, Andersen SJ. DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence? Mol Microbiol 1998; 29:1331–1339.

20

寄生虫基因组学

Malcolm J. Gardner

引言

寄生虫是一类变化多样的生物群体，包括从原生动物疟原虫（*Plasmodium*），到血吸虫病的病原——多细胞蠕虫血吸虫（*Schistosoma*）。它们引起人类很多疾病，在热带和亚热带地区尤为猖獗，这些疾病对发展中国家的经济发展影响巨大：据估计，从1965～1990年，仅疟疾的影响就使得疟疾流行的发展中国家与经济条件相似而没有疟疾的国家相比，国内生产总值（GDP）约低一半[1]。这些疾病消耗的社会资源更难以估量，如学校入学率和科研实力的下降。近年来，寄生虫病对健康和经济发展带来的广泛负面影响被普遍认识，因而激发努力开发有效的对策，如开发新药物、新疫苗和新控制方法。

寄生虫在发达国家和发展中国家也感染许多畜禽，如艾美球虫（coccidian parasite, *Eimeria*），感染鸡和其他禽类。小泰勒虫（*Theileria parva*）是疟原虫的近亲，感染牛，在亚撒哈拉地区的大部分非洲国家，可导致一种致死性疾病——东海岸热病（east coast fever, ECF），该病是非洲农业发展的主要障碍之一，对小农场主尤为棘手，它和其他许多寄生虫一起，降低了农业生产力和产量，使食物短缺。

许多寄生虫在实验室很难进行研究，它们不像那些所谓模式生物，如酵母、果蝇或秀丽新小杆线虫（*Caenorhabditis elegans*）那样，易于在实验室条件下生长、世代时间短、易于遗传操作。许多寄生虫，如疟原虫，有复杂生活史，涉及脊椎动物和无脊椎动物寄主，而且在实验室条件下培养非常复杂、昂贵，或根本不可能。在感染人类的四种疟原虫中，恶性疟原虫（*Plasmodium falciparum*）是唯一能在体外连续培养，但也仅限于在红细胞内期（简称红内期）。所有其他啮齿类和灵长类疟原虫都必须在动物活体中培养，尽管最近有些报道认为伯氏疟原虫（*Plasmodium berghei*，啮齿类疟原虫）的某些有性期可以在体外培养[2]，以及诺氏疟原虫（*Plasmodium knowlesi*，灵长类疟原虫）的红内期可以在培养基中生长[3]，但所有其他人类和啮齿类疟原虫的红内期都必须在动物活体中培养。

这些固有困难使大部分寄生虫研究滞后于其他模式真核生物，然而基因组学却已开始为寄生虫研究者铺平道路。早期基因组学研究是通过对随机挑选cDNA进行测序获得表达序列标签（expressed sequence tag，EST），从而加速了新基因的发现。表达序列标签提供部分或全部基因序列信息，对许多寄生虫它们还能反映出生活史不同时期基因表达模式上的差异，许多这样的早期项目最初是在世界卫生组织热带疾病研究计划（WHO/TDR；http://www.who.int/tdr/）的协调下进行的[4]。许多寄生虫基因组与哺乳动物基因组相比，具有基因密度高、非编码序列较少的特点，因而对随机选取的基因

组片段［基因组调查序列（genome survey sequences，GSS）］进行测序也有助于基因的发现。

随着测序和基因组数据分析技术的进步和测序费用下降，开始对一些寄生虫的一条染色体乃至全基因组进行测序，这些项目快到开花结果的时候了。寄生虫基因组序列信息的大量获得是鼓舞人心的，它们提供了许多新的研究起点，也吸引了更多研究者进入这个领域。功能基因组学研究，主要基于微阵列[5~10]和蛋白质组学手段[11~13]，这给基因组序列增添了基因和蛋白质的表达信息，从而开始对这些寄生虫的生理机制以及它们与其寄主之间的相互作用有了新认识。

本章回顾了对许多重要人类和动物寄生虫开展的基因组测序情况，重点在基因组测序（而不是 EST 计划）和那些已经开始产生效益的寄生虫基因组测序项目，关于寄生虫和其他真核生物的大部分 EST 和基因组测序计划列出了一个表，可在国家生物技术信息中心网站上查询(http://www.ncbi.nlm.nih.gov/PMGifs/ Genomes/EG _ T.html)。

寄生虫基因组计划

寄生虫基因组在大小、结构、基因密度、染色体数目、倍性、碱基组成、重复序列性质以及其他因素上差别极大，例如，原生动物疟原虫（*Plasmodium* spp）是单倍体生物，核基因组小于 25Mb，有 14 条线状染色体，而多细胞生物曼氏血吸虫（*Schistosoma mansoni*）是二倍体，基因组大小约 270Mb。

测序策略必须与待测的基因组和物种相适应，测序策略的选择，受技术条件，如待测物种 DNA 在大肠杆菌中的稳定性、有无大片段文库、计划起始时测序和基因组组装技术能力等的影响，也受经费的多少、该生物对人类和动物健康的重要程度等战略因素的影响。幸运的是，由于仪器改进、凝胶测序仪普遍转为毛细管测序仪、自动化程度提高和实验室流程改进，使得近年来测序费用大幅下降。例如在基因组研究所（The Institute of Genome Research，TIGR），每个碱基的直接费用从 1996 年的 7.8 美元下降到少于 1 美元，而且阅读长度和成功率大大提高，这使得任何一个基因组测序都不再那么昂贵，可以节约资源来对其他基因组测序（如尤氏疟原虫 *Plasmodium yoelii yoelii* 和间日疟原虫 *Plasmodium vivax*，后文将分别讨论），或者可以对那些几年前成本还过于昂贵的大基因组进行测序。随着新测序技术的发展，测序费用还会进一步降低，将来，可以对任何一种感兴趣的寄生虫以及重要寄生虫的多个品系进行全基因组测序。

顶复门

恶性疟原虫

恶性疟原虫（*Plasmodium falciparum*）每年造成约 1～3 百万人死亡，最近，在威康基金桑格研究所（Wellcome Trust Sang er Institute）、TIGR 和斯坦福大学的共同努力下，已完成其全基因组序列[14~19]。该项目始于 1996 年[20]，当时的资助机构和测序中心成立了一个国际联盟，对恶性疟原虫 3D7 株（1976 年分离，能人工培养完成其全部生活史）[21]进行了逐条染色体鸟枪法基因组测序。测序策略的选择适合于当时的测序技

术、序列组装和接缝（closure）技术。

与至今已完成测序的微生物基因组相比，恶性疟原虫最难完成，主要因其核基因组的A+T含量极高（约80%），确定基因组测序需要对标准鸟枪法的几乎每一个部分都作一定修改[17~19]，例如，构建鸟枪法文库的方案需要修改，以防止富含A+T的DNA在凝胶纯化中变性；转座子插入技术用在含有长A和长T重复碱基区域插入转座子，以便准确测定这一段的序列；HAPPY作图技术[16]和光学限制性作图技术[22,23]用于对克隆重叠群（contig）排序和确认最终染色体序列。此外，还开发了两个新基因搜寻软件，因为已有的基因搜寻软件对预测高基因密度真核基因组中的基因不是最优化的[24,25]。完成全基因组序列花了6年多时间，难度主要在封闭序列中几百个富含AT的缺口，公布的基因组序列还有约93个缺口（大部分缺口小于2500个碱基），封闭缺口的工作仍在继续进行。

恶性疟原虫基因组约23Mb，编码约5300个基因[14]，54%的基因含有内含子。总之，恶性疟原虫基因编码区比其他已测序的真核生物要长，如裂殖酵母（前者为2.3kb，后者为1.4kb），还不清楚基因长度增加的原因，60%以上的编码蛋白质与其他生物相似性很低或无，功能也未知。恶性疟原虫中编码这些假定蛋白质的基因比例比其他已测序的物种要高，这可能是疟原虫与其他研究较深入真核生物之间进化距离比较远的缘故。

对预测蛋白质组的分析可为疟原虫的代谢和物质转运提供一个轮廓，然而，由于缺乏代谢途径中的一些酶和酶亚基，寄生虫代谢的一些特点仍然不清楚；况且一些酶的亚细胞定位与其他生物的已知定位不同，使得很难对这种寄生虫的代谢过程准确重构。不过，分析基因组序列对了解疟原虫的生物化学仍可提供有价值的思考，并可提出尚需在实验室进一步研究的问题（参见第6章）。

恶性疟原虫的代谢能力、有机营养和离子转运能力比非寄生生物低，与已知的酶序列进行相似性比较发现，只有14%蛋白质编码酶，这个比例与其他已测序真核生物相比是相当低的。与此类似的是，恶性疟原虫基因组与其他非寄生真核微生物，如酿酒酵母（*Saccharomyces cerevisiae*）和裂殖酵母相比，编码相对小的膜转运体系统，早期的生化研究暗示，红内期寄生虫ATP的产生需要依赖糖酵解，恶性疟原虫基因组编码了从葡萄糖-6-磷酸到丙酮酸的糖酵解途径全部酶类，以及将丙酮酸转化为乳酸的代谢酶类。

三羧酸循环（TCA）的全部酶类都已确定，但TCA的功能还不清楚。丙酮酸脱氢酶，一般定位在线粒体中，可将丙酮酸转化为乙酰辅酶A，后者是TCA循环第一步所必需的；而在恶性疟原虫中，预测它定位在顶端体（apicoplast，一种残遗的质体）中。此外，苹果酸脱氢酶在其他真核细胞中定位在线粒体内，而它却存在疟原虫细胞质内，而且在TCA循环中，该酶被线粒体中存在的苹果酸-醌氧化还原酶所代替。这些特点表明，至少在红内期，TCA循环提供血红素生物合成等合成途径的中间体，而不是对糖酵解产物进行完全氧化。

TCA循环在生活史其他阶段中的作用还不明确，但是蛋白质组学研究揭示，一些TCA循环的酶类在配子体内比在红内期寄生虫中含量更丰富，表明TCA循环对在蚊子体内进行的有性阶段更为重要[11,12]。意外的是，ATP合成酶中Fo a、b亚基缺失，这

意味着 ATP 合成酶在疟原虫中可能是无功能的，也可能因为编码这些亚基的基因较短，位于基因组未测序区域而没有被识别，或许这两个亚基序列改变很大，以至于不容易通过序列相似性方法识别出来。

从基因组序列推导出恶性疟原虫代谢不寻常的特点是，缺乏糖异生作用（gluconeogenesis），缺乏合成氨基酸的任何酶（除了含有一些氨基酸互相转换所需的酶之外）。氨基酸生物合成途径的缺失和已知氨基酸转运体的同源物表观缺失，突出反映了寄生虫在氨基酸需求上对寄主的依赖性，至少在红内期是这样的：氨基酸通过消化食物液泡中的血红蛋白获得，尚不知道寄生虫是如何在生活史的蚊子体内期获得氨基酸。脂肪酸和类异戊二烯的生物合成发生在顶端体，这两个途径与植物和细菌类相似，而与动物不同，这为研制新抗疟药物提供了很多潜在的靶标[26~28]，总之，疟原虫的代谢和转运能力比那些非寄生生物弱，这与它们营寄生生活方式一致（参见第 6 章图 3）。

恶性疟原虫与其他已测序真核生物基因组的另一个明显不同之处，在于前者含有大量参与免疫逃避和其他与寄主-寄生虫相互作用有关的基因（分别占总基因 3.9% 和 1.3%）。3D7 株基因组含 59 个 *var* 基因，可编码高度多形性蛋白——恶性疟原虫红细胞膜蛋白 1（*P. falciparum* erythrocyte membrane protein 1，PfEMP1）。这些蛋白质在受染红细胞表面表达，介导受染红细胞与寄主毛细血管内皮细胞之间的细胞黏附，随之引起受染细胞在许多器官中的滞留，包括脑部。PfEMP1 蛋白是保护性抗体反应的靶抗原，但不同 *var* 基因转录导致抗原变异，因而逃脱了寄主的免疫攻击。在 3D7 基因组中有 149 个 *rif* 基因，编码另一组蛋白质，称为 rifin。它们也是在感染红细胞表面表达的蛋白，也呈现抗原变异，尚不清楚它们的确切功能。第三组蛋白为 STEVOR（在 3D7 基因组中有 28 个成员），与 rifin 序列相似。

PfEMP1、rifin 和 STEVOR 家族成员呈现出序列的多样性，它们的编码基因成簇分布，大多数定位在亚端粒区，伴随着几种重复序列，重复序列的存在为这些高度多形性蛋白等位基因之间的重组提供了方便，因而，有助于抗原多样性的产生。

恶性疟原虫 3D7 已经培养了多年，而没有面临过免疫压力，培养的寄生虫可能发生染色体断裂和融合，这将导致亚端粒区的缺失和基因丢失，3D7 的几条染色体似乎已发生过这样的短截。将 3D7 与近年临床分离的品系进行基因组比较，可得知 3D7 中涉及免疫逃避和寄主-寄生虫相互作用的基因是否在野生型寄生虫中具有代表性，这项研究将在近期开展（N.Hall，桑格研究所，个人通讯，2003）。

恶性疟原虫基因组计划是第一个完成的重要人类寄生虫基因组计划，该计划的重要性在于四个私立和公众支持机构组织每半年一度的会议，使疟疾研究团体能在项目中广泛沟通。这些会议讨论项目进展和测序者遇到的技术难题，以求最终完成基因组全序列，并计划将来的研究活动，如功能基因组学和蛋白质组学研究，以及数据发布制度。

尽管 1997 年通过的数据发布制度并不能使每个人满意[29,30]，但它的确建立了一个流程，向研究团体发布初期序列和注解，以“使生物学实验跳跃式启动（jump start）”，也迎合了测序中心对基因组序列进行组装，注释和公布的要求，制度建立后，有几十篇研究论文发表，并在文中对得以使用初期序列数据表示感谢。

在这些会议中，详细讨论的另一个议题是建立一个集中的、用户友好的数据库，以方便疟疾研究者获得恶性疟原虫基因组序列信息和序列分析工具，而不再访问其他网

站。讨论的结果于2000年6月在宾夕法尼亚大学建立了PlasmoDB[31,32]，它后来扩展到包括啮齿类、灵长类和人类疟原虫及其他种的序列信息，还包含了EST[33]、SAGE[34,35]和微阵列[7,8]数据有关基因表达的信息。

间日疟原虫

间日疟原虫（Plasmodium vivax）在人类疟疾寄生虫中的重要性排名第二，它每年引起7000～8000万人类疟疾[36]，在亚洲、大洋洲、美洲和非洲的部分地区流行，是非洲大陆以外半数疟疾患者的病原。间日疟原虫对来自无疟疾国家的旅行者和军队是一个严重威胁，尽管很少致死，但间日疟能使人衰弱无力，因而削弱生活质量和生产力。在几个国家已发现间日疟原虫对氯喹的抗药性[37～39]，这将导致全球范围内间日疟病例的增加和流行。

间日疟原虫和恶性疟原虫有许多不同之处，间日疟原虫在肝脏中为静息状态，称为休眠体（hypnozoite），在最初感染很长时间后才发生激活和随后的发展，并引起疾病复发。在红内期，间日疟原虫侵袭网织红细胞，而不是成熟红细胞，使寄生虫血症比恶性疟原虫感染要轻。另外，间日疟原虫侵袭网织红细胞需要Duffy抗原和趋化因子受体之间的相互作用[40]，后者在网织红细胞表面表达，而Duffy结合蛋白在裂殖子表面表达[41]。间日疟原虫与恶性疟原虫的另一个不同在于，前者在毛细血管中不被滞留，而滞留牵涉到脑型疟疾及恶性疟原虫引起的其他严重症状。

尽管间日疟原虫是人类的重要病原，但对它的研究比恶性疟原虫少，不像恶性疟原虫那样[42,43]，尚不能大规模体外培养红内期间日疟原虫。因此，间日疟原虫的实验室培养必须用灵长类动物，这在大多数实验室因太昂贵而无法实现。用恶性疟原虫基因组计划剩余的经费，TIGR和海军医学研究中心（Naval Medical Research Center）正在对间日疟原虫基因组实行全基因组鸟枪法测序。

间日疟原虫核基因组由12～14条线状染色体组成，大小在1.2～3.5Mb[44]。早期估计基因组大小在35～40Mb，但根据其他疟原虫的基因组和初步测序结果判断，早期的估计值偏大，应在23～25Mb。对间日疟原虫基因组的大规模取样（从绿豆核酸酶消化构建的基因组DNA文库中选取11 000个考察序列），已对它的编码潜力多了一些理解[46]，已知疟原虫基因同源的基因已被确定，间日疟原虫与恶性疟原虫蛋白质组也含有相似的蛋白质数目，间日疟原虫中一个大的、包括600～1000个变体蛋白、命名为*vir*的多基因家族最近被确定[47]，在几种啮齿类疟原虫中发现了直系同源基因[45,48]，而恶性疟原虫却没有。

基因作图研究[45,49,50]、小规模测序计划[47,51]和基因组范围的共线性图谱（synteny map）[45]显示，疟原虫各种之间具有广泛保守的基因共线性，基因种类和顺序在几百kb范围内显示出高度保守，例如，间日疟原虫基因组中一个200kb的片段包含36个连续基因，基因的顺序、方向和结构与恶性疟原虫3号染色体上的直系同源片段完全一致[51]，疟原虫各种之间共线性的程度，随着它们之间种系发生距离的增加而减少，这一点在其他生物也如此。

从萨尔瓦多一个自然感染患者体内分离到的间日疟原虫Salvador I品系[52]，正在TIGR用全基因组鸟枪法进行随机测序，构建了小片段和中长度片段基因组文库，并已

完成9倍覆盖率测序[53]。在基因组组装完成后，将初步获得克隆重叠群中的预测可读框（ORF）与已知蛋白数据库进行比较，揭示出37%ORF与已知蛋白有序列相似性，其中78%与恶性疟原虫蛋白质同源，其余63%的ORF与任何已知蛋白没有相似性，这与恶性疟原虫基因组分析得到的比例相似[14]，缺口封闭正在进行，预计2004年将完成整个基因组序列。幸运的是，缺口封闭的初步工作显示，完成间日疟原虫基因组不会像恶性疟原虫那样困难，因为前者基因组DNA中的GC含量较高（Jane Carlton，个人通讯，2003）。

尤氏疟原虫和其他啮齿类疟疾寄生虫

尤氏疟原虫（*Plasmodium yoelii*）是一种啮齿类疟疾寄生虫，在许多实验室作为动物模型，广泛用于研究在子孢子和红外期的抗原表达，并用来检测这些抗原的疫苗。用恶性疟原虫基因组计划剩余的经费，TIGR和海军医学研究中心对尤氏疟原虫基因组进行了5倍覆盖率的测序，并将二者的基因组进行了比较分析[45]，尤氏疟原虫基因组为23Mb，编码约5900个基因，这个数字比恶性疟原虫多一些，这是因为尤氏疟原虫中多了编码变异抗原的838个小基因（*yir*基因）。尤氏疟原虫有一半预测蛋白质与恶性疟原虫直系同源，这些蛋白质几乎都定位在染色体的中部区域，而那些特异性基因，如尤氏疟原虫*yir*基因和恶性疟原虫*rif*，*stevor*和*var*基因都定位在染色体亚端粒区域。

早先，用脉冲凝胶电泳技术揭示了啮齿类和人类疟原虫基因共线性区域的广泛性，恶性疟原虫和尤氏疟原虫基因组序列的获得，使基因共线性研究可以借助生物信息学技术来完成，这显然要精确得多。第一步，未排序的尤氏疟原虫克隆重叠群和恶性疟原虫基因组都以6种阅读框来翻译，然后，两个物种间所有精确匹配长度超过5个氨基酸的序列都被输入最小唯一匹配（minimal unique matches，MUMmer）程序进行计算分析。用这种方法，尤氏疟原虫70%克隆重叠群可以在恶性疟原虫基因组中找到位置，表明这两种疟原虫之间有广泛的基因排列顺序保守性，尤氏疟原虫中大部分保守克隆重叠群都被标示在恶性疟原虫染色体的中部区域。

为了辨认大片段基因共线性区域以及共线性断裂点，尤氏疟原虫保守克隆重叠群之间的连接部分通过配对信息（mate-pair information）、鉴定跨越克隆重叠群缺口EST以及PCR扩增相邻克隆重叠群片段之间的序列等方式被确定，用这些研究确定了457个相连克隆重叠群，共长800kb，随后，通过染色体特异性标记定位在尤氏疟原虫染色体上。将相连尤氏疟原虫克隆重叠群作图到恶性疟原虫的基因组中，提供了二者基因共线性图谱，并确定了共线性断裂位点[45]。尤氏疟原虫70%基因与恶性疟原虫假定直系基因有相同的排列顺序和方向。尤氏疟原虫基因组中共线性断裂点往往伴随编码rRNA的基因位点，表明染色体在rRNA位点处的断裂和重组导致共线性断裂。在标注出共线性区域的同时，分析共线性区域内直系同源基因的结构，揭示出基因中外显子结构的高度保守，因此，正如所指出的那样，疟原虫中直系同源基因的排列顺序非常有助于阐明基因结构[54,55]。

尽管这种中度覆盖率测序可以提供大量有关尤氏疟原虫基因组以及它与恶性疟原虫基因组的相同和不同之处的有用信息，序列数据的不完整仍妨碍了更全面的分析。首

先，草图还未经校订，缺口封闭经常能发现和纠正由重复序列引起的序列拼接错误，所以与那些已经“完成”了的基因组测序相比，尤氏疟原虫基因组序列中可能还有更多错误和不确定性。其次，尤氏疟原虫的许多基因模型还不完善。第三，缺乏完整的染色体序列，对尤氏疟原虫和恶性疟原虫进行基因共线性的保守分析非常困难。另一方面，低度和中度覆盖率数据对疟原虫这样基因密度高的生物非常有用，它为许多实验研究者提供了方便。

另两种广泛研究的啮齿类疟原虫——伯氏疟原虫（*Plasmodium berghei*）和夏氏疟原虫（*P. chabaudi*）已获得5倍覆盖率测序的基因组序列，这可以确定90%以上的基因。桑格研究所在对灵长类疟原虫——诺氏疟原虫（*P. knowlesi*）进行基因组测序，诺氏疟原虫与人类的间日疟原虫亲缘关系很近，常被用来研究侵袭红细胞的寄生虫分子，以及检测疫苗候选抗原和疫苗给药系统[56]。一般是在恒河猴中生长的诺氏疟原虫，已研发出可以体外长期培养（最长18个月）的技术，在体外长期培养的寄生虫又可重新适应在体内生长。另外，诺氏疟原虫的体外转染技术和快速基因敲除（gene knock-out）技术已经可行，诺氏疟原虫模型的这些特性和从其基因组测序收集到的信息以及与其他疟原虫的序列比较，使得对寄生虫基因功能和寄主-寄生虫相互作用的研究，可以在一个与人类病原非常近似的系统中开展。

梨浆虫

梨浆虫（*Piroplasms*）是一类蜱传播寄生虫，可以侵袭红细胞，有时也侵袭哺乳动物寄主的其他细胞。不像它们的近亲疟原虫，梨浆虫不会从血红蛋白生成疟色素。最重要的两种梨浆虫是巴贝虫（*Babesia*）和泰勒虫（*Theileria*），它们是动物寄生虫，不过巴贝虫偶尔也感染老年、免疫缺陷和切除脾脏的人。

牛巴贝虫（*Babesia bovis*）和二联巴贝虫（*Babesia bigemina*）是热带和亚热带地区牛巴贝虫病（蜱热）的元凶，分歧巴贝虫（*Babesia divergens*）出现于温带。巴贝虫在被感染牛中引起贫血和热病，导致牛生产力下降，并可诱导怀孕动物流产。对巴贝虫的预防通常是通过活减毒疫苗，或者对牛施用杀蜱剂，以防止蜱的侵袭和寄生虫传播，这些措施花费昂贵，即使是在高度发达国家也难以长期维持。

牛巴贝虫的基因组通过脉冲凝胶电泳已了解到包括四条染色体，长度在1.4～3.2Mb，共计9.4Mb[57]，全基因组鸟枪法测序可望在近期开展，在从被感染红细胞获得的环状DNA文库中也开始产生EST序列(http://www.sanger.ac.uk/Projects/B-bovis/)。

泰勒寄生虫广泛感染家养和野生的各种反刍动物，包括奶牛、绵羊、水牛和鹿。在经济上最重要的两种是小泰勒虫（*Theileria parva*）和环形泰勒虫（*Theileria annulata*）。小泰勒虫是牛东海岸热病（east coast fever，ECF）的病原，该病导致非洲亚撒哈拉地区家畜高致病率和死亡率[58]，每年东海岸热病致死100万头牛，经济损失预计达16 800万美元以上[59]。饲料中混有带病原的褐色耳蜱（*Rhipicephalus appendiculatus*）可以使子孢子开始感染寄主，子孢子侵袭寄主淋巴细胞，然后发展为多核的裂殖体。寄生虫在淋巴细胞细胞质中出现后，诱导感染细胞恶性转化，裂殖体感染的淋巴细胞迅速增生扩散，使感染的动物在一个月内死于白血病样的疾病。感染了小泰勒虫的转化细胞

可以在体外无限增生，并可对裸鼠致瘤，但转化现象可以用杀寄生虫药物来逆转，因此，泰勒虫为研究细胞转化的诱导、维持和逆转提供了一种独一无二的模型[60]，研究小泰勒虫转化哺乳动物细胞的机制，有助于揭示人类癌症的相关现象。

与巴贝虫的防治类似，对小泰勒虫的主要控制方法也是使用减毒活疫苗，以及对牛施用杀蜱剂，这些方法在非洲不发达地区都是很难维持和支付的。开发蛋白亚基疫苗，可能是潜在最有效的 ECF 控制方法，也是目前主要的研究目标。研究使用活减毒疫苗在体内引起的免疫应答，发现保护机制是由 $CD8^+$ 杀伤裂殖子感染细胞引起，$CD8^+$ 是主要的组织相容性限制的 T 细胞。

过去几年，国际家畜研究所（International Livestock Research Institute，ILRI，内罗毕，肯尼亚）一直试图确定在裂殖子感染细胞表面表达、成为保护性 $CD8^+$ T 细胞的靶标的寄生虫抗原，却未能成功，主要原因是寄生虫的孢内生活方式。ILRI 和 TIGR 合作对小泰勒虫的基因组测序，以帮助寻找在裂殖子感染细胞表达的抗原，测序采用全基因组鸟枪法战略。寄生虫的全部四条染色体已经完成，尽管 3 号染色体中一段约 120kb 的、含有高度重复 Tpr 序列的区域还不能完全破译。全基因组约 8.4Mb，正在开展对基因组的注解，小泰勒虫新基因的初步序列，已被 ILRI 用来寻找在裂殖子感染的细胞表面表达的抗原，用微阵列研究寄生虫整个生活史中基因表达的变化也在计划之中。

环形泰勒虫在地中海周边国家和亚洲流行，像小泰勒虫一样，它也感染牛，引起致命的细胞增殖紊乱，但它转化巨噬细胞，而不是淋巴细胞[60]，因此，小泰勒虫和环形泰勒虫有许多共同点，可感染寄主不同细胞。环形泰勒虫的基因组正在由桑格研究所测序，将小泰勒虫和环形泰勒虫的基因组进行比较，以确定物种特异性基因和保守基因，将会找到这两种寄生虫侵袭和感染不同寄主细胞的机制。

刚地弓形虫

刚地弓形虫（*Toxoplasma gondii*）是一种普遍存在的寄生虫，可以感染多种动物。刚地弓形虫是 HIV/AIDS（human immunodeficiency virus，人类免疫缺陷病毒/acquired immodeficiency syndrom，获得性免疫缺陷综合征）患者或其他免疫缺陷患者中常见的机会病原体，它可在子宫内感染胚胎，从而引起儿童严重先天缺陷，它也是一种重要性畜病原。由于它对人畜的致病性，且易于在实验室（体外或动物体内）培养，因此被广泛研究，从而成为研究其亲缘寄生虫（如疟原虫等顶复门）时的一种便于操作的模式生物。已经可以构建和分离弓形虫突变体，实现遗传交换，用外源 DNA 稳定或瞬时转化，制备基因敲除和等位基因置换以及实现插入突变（综述参见文献［61］)。

刚地弓形虫的单倍体基因组包括 11 条染色体，长度从 2～10Mb 以上，总共约 80Mb[62]，同疟原虫一样，刚地弓形虫也有一个 6kb 线粒体基因组和一个 35kb 顶端体基因组[63]，两个平行基因组测序计划正在进行（表 1)。一个计划由 TIGR 和宾夕法尼亚大学合作进行，利用鸟枪法战略来决定最常见于 HIV/AIDS 病人的一个品系（Ⅱ型）的基因组序列，开始计划的目的是对基因组进行 3 倍覆盖率测序，后来因为测序费用降低和可用资金增加，计划扩展到 8 倍测序并进行基因组注解，计划目前已完成 7 倍覆盖率，初步克隆重叠群已经发布。第二个计划由桑格研究所执行（表 1），目标是对 1 号染色体测序，并从一个基因组细菌人工染色体（bacterial artificial chromosome，BAC）

文库中产生末端序列。

在对刚地弓形虫基因组进行注解时，有一个资源将会非常有用，那就是虫体生活史中来自不同时期 EST 的庞大数据库（目前在 GenBank 中已经达到约 64 000 条 EST)[64~66]，EST 已经成簇排列[65,67]，并用来制作基因索引(http://www.tigr.org/tdb/tgi/tggi)。这些拼接好的 EST，可为基因搜寻软件提供大量训练数据，以便提供更加准确和完整的基因预测；在当时注释疟原虫基因组时，只能用少量通过实验确认的基因序列来训练基因搜寻软件。一个载于疟原虫数据库（PlasmoDB）的弓形虫数据库(ToxoDB)（http：//toxodb.org/ToxoDB.shtml）已经建立，可提供基因组序列的有关信息[68]。

表 1　部分寄生虫基因组项目[a]

物种	疾病	基因组特征
顶复虫亚门(Apicomplexa)		
恶性疟原虫(*Plasmodium falciparum*)	疟疾(人类)	23Mb，20%G+C
间日疟原虫(*Plasmodium vivax*)	疟疾(人类)	25Mb，38% G+C
尤氏疟原虫(*Plasmodium yoelii*)	疟疾(啮齿类)	23Mb，23% G+C
伯氏疟原虫(*Plasmodium berghei*)	疟疾(啮齿类)	25Mb，20% G+C
夏氏疟原虫(*Plasmodium chabaudi*)	疟疾(啮齿类)	25Mb，20% G+C
文氏疟原虫(*Plasmodium vinckei*)	疟疾(啮齿类)	25Mb，20% G+C
诺氏疟原虫(*Plasmodium knowlesi*)	疟疾(灵长类)	25Mb，38% G+C
鸡疟原虫(*Plasmodium gallinaceum*)	疟疾(鸡)	25Mb，20% G+C
小泰勒虫(*Theileria parva*)	东海岸热(牛)	8.5Mb，34% G+C
环形泰勒虫(*Theileria annulata*)	泰勒虫病(牛)	8.5Mb，32.5% G+C
刚地弓形虫(*Toxoplasma gondii*)	弓形体病	80Mb，52% G+C
小隐孢子虫(*Cryptosporidium parvum*)	痢疾	9.4Mb，68% G+C
动基体目(Kinetoplastida)		
巨大利什曼原虫(*Leishmania major*)	利什曼病	34MB，63%G+C
布氏锥虫(*Trypanosoma brucei*)	非洲昏睡病	35Mb，50%G+C
克氏锥虫(*Trypanosoma cruzi*)	查格斯病	40Mb，50%G+C
微孢子虫目(Microsporidia)		
家兔脑胞内原虫(*Encephalitozoon cuniculi*)	胃肠感染	2.9Mb，47%G+C
六鞭虫科(Hexamitidae)		
兰氏贾第鞭毛虫(*Giardia lamblia*)	贾第鞭毛虫病	12Mb，46%G+C
内阿米巴虫科(Entamoebidae)		
溶组织内阿米巴(*Entamoeba histolytica*)	阿米巴性痢疾	20Mb，24%G+C
吸虫纲(Trematoda)		
曼氏血吸虫(*Schistosoma mansoni*)	比哈西亚症(Bilharzia)	270Mb，37%G+C
线虫纲(Nematoda)		
马来布鲁线虫(*Brugia malayi*)	淋巴腺丝虫病	110Mb，71%G+C

a 仅指基因组测序项目；不包括 EST 计划。关于真核基因组、EST 和 GSS 计划简表，可从美国国家生物技术信息中心网站获得(http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/EG_T.html)。参见文中所指参考文献。

小隐孢子虫

小隐孢子虫（*Cryptosporidium parvum*）引起人和动物急性胃肠疾病。在许多发展中国家，它是传播最广泛的肠道病原之一，在发达国家，它是一种新兴起（emerging）的病原，它在免疫缺陷个体中流行[69]。尽管医学上重要，对小隐孢子虫的研究还是很少，主要因为它不能在体外有效地培养。小隐孢子虫的染色体组型由脉冲凝胶电泳确定，包含8条染色体，长度从0.94～1.44Mb不等，总共约9.4Mb[70,71]。基因组测序始于EST和GSS策略[72～74]，最近又有两个对小隐孢子虫全基因组测序计划开始启动（表1），其中一个是针对基因型Ⅰ分离物测序，由它引起大多数人类感染；另一个是针对基因型Ⅱ分离物测序，它引起所有动物感染和一些人类感染。

不像至今研究过的其他顶复门寄生虫，小隐孢子虫不含有顶端体基因组[75]，可以推测，小隐孢子虫在进化中的某个阶段“丢失”了顶端体。如果事实确如此，那么将小隐孢子虫的基因组与其他顶复门虫，尤其是恶性疟原虫的基因组进行比较将非常有意思。顶端体在其他顶复门虫体中起关键的生化作用，在恶性疟原虫中，10%以上预测的核基因都编码着定位到顶端体的蛋白质[14,76]。

动基体目

动基体目（Kinetoplastida）是一大类有鞭毛的原生动物，可在很多不同生物体内引起疾病，包括植物和动物。在人体中，动基体目原生动物引起的疾病包括非洲昏睡病、查格斯病（Chagas disease）、皮肤和内脏利什曼病，每年有几亿人受到这些疾病的折磨。除了作为病原非常重要外，动基体目寄生虫还展示出非常有趣的生物学特性，如抗原变异、动基体、反式剪接（*trans*-splicing）、RNA编辑以及糖基磷脂酰肌醇锚定蛋白。不像许多其他寄生虫，动基体虫较容易在实验室进行研究，这个优点加上它们是重要病原物，以及它们奇特的生物学特性，使得它们成为非常流行的模式系统。几种动基体寄生虫正在测序，这将为比较基因组学提供丰富的数据。

巨大利什曼原虫

巨大利什曼原虫（*Leishmania major*）由白蛉传播，它侵袭和寄居在寄主巨噬细胞内，引起皮肤和内脏利什曼病。它是二倍体，核基因组约34Mb，包括36对染色体，富含GC（～63%），基因组含重复序列，包括六碱基端粒重复（hexameric telomeric repeats），简单序列重复和可转座元件。利什曼原虫基因组网络（http://www.sanger.ac.uk/Projects/L_major/），由WHO/TDR资助，正在组织巨大利什曼原虫的基因组测序。由于基因组结构的复杂性，采用了多种测序策略，包括对重叠黏粒克隆测序[77]，对脉冲凝胶分离染色体进行鸟枪法测序和BAC克隆末端测序。

最小1号染色体（269kb）的测序已经完成[78]，该染色体含79个编码蛋白质基因，定位在中部257kb区域，该区域的侧翼分别是4kb和8kb非编码区，然后是亚端粒和端粒重复序列。该染色体的有趣特点是蛋白编码基因的排列，前29个基因覆盖73kb，定位于DNA的一条链，其他50个基因（182kb）却定位在另一条链，形成两个多顺反子

基因簇的头对头地排列，中间由 1.6kb 区域（转换区域）相隔，转换区域在转录起始时起作用，作为复制原点或着丝粒。当切除 1 号染色体转换区域似乎并不影响插入在其附近转基因的转录，也不改变染色体在有丝分裂时的稳定性，但是，如果删除 1 号染色体所有拷贝中的转换区域，则在所有被分析的克隆中都多了一份拷贝的染色体[79]，所以，转换区域与基因表达或染色体稳定性无关，但对寄生虫的存活是必须的。许多染色体已完成测序，其余染色体的测序也正在进行(http://www.sanger.ac.uk/Projects/L_major/progress.shtml)，如同在 1 号染色体中发现的一样，大型多顺反子基因簇是巨大利什曼原虫基因组的普遍特点。

布氏锥虫

人类的非洲昏睡病是由于感染了布氏冈比亚锥虫（*Trypanosoma brucei gambiense*）和布氏罗德西亚锥虫（*Trypanosoma brucei rhodesiense*）后引起的，在非洲亚撒哈拉广大地区由采采蝇传播。发现采采蝇的地区有 5 亿以上人口生活，估计每年有几十万昏睡病病例，引起约15 000人死亡。还没有预防感染疫苗，而现有药物大多有毒性或价格昂贵。第三个种，布氏布氏锥虫（*Trypanosoma brucei brucei*）引起家畜的致死性疾病——那加那病（Nagama），非洲大部分地区不能饲养对这种感染性疾病易感的动物，从而阻碍了这些地区肉类和其他产品的生产发展。

布氏锥虫基因组测序由 TIGR 和桑格研究所进行（表 1），基因组的组织很复杂，至少有 11 对 Mb 大小的染色体（约 1～6Mb），许多中等大小的染色体（0.2～0.9Mb），还有 50～100 个线性小染色体（0.05～0.15Mb）。对 BAC 文库逐个测序，边测序边定位，基因组的重复性非常麻烦，到截稿时，所有染色体序列都已完成并注解[80,81]。与在巨大利什曼原虫中发现的情况类似，布氏锥虫的先注解的两条染色体上，基因也是排列成长的多顺反子单向基因簇，有的长达 250kb，含有 100 个以上的基因，基因簇之间的 DNA 链转换，伴随着 G+C 偏衡（G+C skew）的统计学变化（一种计算两条 DNA 链间碱基组成不对称性的方法），这表明观察到的碱基组成不对称性变化可能与转录有关，或与转录耦联修复（transcription-coupled repair）有关[80]。其他特点包括：新基因家族、有的基因家族成员串联排列，有大量反转录元件（retroelement）和染色体片段重复，这些使得有条件研究表面变异糖蛋白基因在亚端粒上的表达位点，而这些糖蛋白基因在抗原变异中发挥作用。

克氏锥虫

克氏锥虫（*Trypanosoma cruzi*）是查格斯病的病原，该病在南美洲和中美洲广大地区流行，每年感染人数 1800 万以上，死亡超过 5 万。虫体在锥蝽（接吻虫）阶段传播，通过叮咬被感染者，然后将感染物以粪便形式传播给其他寄主，含有寄生虫的粪便通过被摩擦进入暴露的伤口或眼睛而传染，一旦进入人类寄主，克氏锥虫就会侵袭很多类型的细胞，急性感染一般引起中度症状，并可发展成为慢性感染，导致器官损伤，尤其是对心脏和消化道。慢性感染患者三分之一以上死于此病。

克氏锥虫基因组由 40 对以上染色体组成，共 44Mb[82]，一个国际性基因组计划，克氏锥虫基因组启动计划（*T. cruzi* Genome Initiative, http://www.dbbm.fiocruz.br/

TcruziDB/)，制备了酵母人工染色体（yeast artificial chromosome，YAC）、BAC、黏粒文库；此外，EST 文库的构建也进行了好几年[83-85]。全基因组计划由乌普萨拉大学、西雅图生物医学研究所和 TIGR 启动（表 1)，起初，计划采用以 BAC 为基础的测序策略，但是基因组的高度重复特性干扰了指纹识别和测序 BAC 的选择，后来，策略改为全基因组鸟枪法测序，目的是获得 20 倍覆盖率的序列，因为被测序的分离物由两个单体型组成，随机测序阶段已经达到 14 倍覆盖率，由于存在大量重复序列，还有两个单体型，对基因组的精确组装也充满了挑战。

微孢子虫目

微孢子虫（Microsporidia）是专性细胞间真核寄生虫，引起人体消化道和神经系统感染，尤其是对免疫缺陷的个体。它们有简单、特别的生活史，包括在寄主细胞内的增殖（卵片生殖）和形成能转播到其他寄主的孢子（孢子生殖)，它们没有线粒体或过氧化物酶体，而它们的种系发生地位也有争议，最初认为是无线粒体的原生动物，以后认为是丢失了线粒体的真菌。

一种微孢子虫——家兔脑胞内原虫（*Encephalitozoon cuniculi*）的基因组序列已于近日完成，其基因组与那些自由生活的真核生物相比有高度致密性[86]。基因组只有 2.9Mb，包括 11 条染色体，编码约 2000 个基因，基因密度约为每 1.25kb 含有一个基因，这是裂殖酵母和酿酒酵母的 4 倍。这么高的基因密度反映出基因间间隔区域很短(平均 129 碱基)，内含子少，编码序列（平均长 1077 个碱基）也比其他已测序的生物短，例如，家兔脑胞内原虫的 350 个蛋白比它们在酿酒酵母中直系同源物平均长度短 15%，这与恶性疟原虫中发现的很不一样，因为恶性疟原虫中编码序列的平均长度(2283 个碱基）比其他已测序真核生物要长得多（1300～1600 碱基)。

对基因组序列的分析，为深入理解家兔脑胞内原虫的生化和细胞生物学提供了很多信息，它似乎缺乏 TCA 循环的酶类、呼吸电子传递系统、氨基酸生物合成、嘌呤和嘧啶的从头合成、脂肪酸合成酶和其他途径，只含有少数膜转运体。这些特点证实该寄生虫在重要生化底物的需求对寄主的依赖性，尽管家兔脑胞内原虫不含可辨认的线粒体或线粒体 DNA，但它含有与酿酒酵母线粒体中 Fe-S 蛋白复合体组成蛋白具有直系同源性的蛋白，因此，这种生物可能含有退化的细胞器，称为线粒剩体（mitosome)，其功能涉及保护细胞免受氧化物的损伤。

六鞭虫科

兰氏贾第鞭毛虫（*Giardia lamblia*）是一种有鞭毛的真核寄生虫，感染人和其他哺乳动物的小肠，引起许多种消化道症状，尤其是腹泻。感染始于摄取休眠孢子，它在小肠内发育为滋养体。贾第鞭毛虫及其亲缘的寄生虫，如毛滴虫（*Trichomomas*）属最原始的真核生物，无线粒体、核仁和过氧化物酶体，但保留了其他真核细胞的典型特征。它是多倍体的，每个滋养体含有两个明显相同或相似的核，每个核有基因组的完整拷贝[87]，基因组大小估计为 12Mb，由 5 条 1.6～3.8Mb 的染色体组成。

海洋生物实验室（The Marine Biology Laboratory）正在用全基因组鸟枪法战略对兰氏贾第鞭毛虫基因组测序[88]，随机测序阶段已获得近 7 倍覆盖率的序列，克隆重叠群代表了至少 90% 基因组（http://jbpc.mbl.edu/Giardia-HTML/）。阴道毛滴虫（*Trichomonas vaginalis*）是引起不孕最常见的性传播疾病病原，由 TIGR 进行基因组测序（Jane Carlton，个人通讯，2003），基因组序列除可方便实验室研究外，还能更好地理解这些原始寄生虫和一些高等真核生物之间的种系发生关系。

内阿米巴虫科

溶组织内阿米巴（*Entamoeba histolytica*）是一种原生寄生虫，引起感染者的痢疾和肝脓肿，由污染寄生虫的水源而传播，每年受感染人数约 4000～5000 万。溶组织内阿米巴的基因组约 20Mb，含有约 14 条线状染色体和环形 DNA 分子，含有丰富的重复序列[89]，由 TIGR 和桑格研究所合作对基因组测序，综合两个测序中心读出的鸟枪片段序列，一起组装基因组序列。

对此基因组测序有一个不常见的问题，就是编码 rRNA 基因存在附加体，附加体的序列占原始鸟枪法文库序列的 15%，用限制酶酶切附加体，再通过电泳去除线性化附加体，可构建含有较少 rRNA 序列的基因组文库[90]，随机测序已达约 9 倍覆盖率（http://www.tigr.org/tdb/e2k1/eha1/； http://www.sanger.ac.uk/Projects/E_histolytica/）。

为了方便比较研究，几种其他内阿米巴虫的部分基因组测序正在进行，包括迪斯帕内阿米巴（*Entamoeba dispar*）——一种感染人类的非入侵性种类，但与溶组织内阿米巴在形态上很难分辨，这有助于确定致病毒力因子。

吸虫纲

裂体吸虫（*Schistosomes*，俗称血吸虫）在非洲、亚洲和美洲发现，通过钉螺释放游动的、侵袭性尾蚴到水中传播，尾蚴在哺乳动物寄主皮肤上打洞，转化为童虫，然后进入循环系统。童虫发育为成对的雄性和雌性成虫体，寄居在寄主的血管并产卵，卵通过寄主的粪便或尿（因血吸虫种类不同而不同）释放到环境中。卵发育为毛蚴，再感染钉螺，从而完成生活史。感染人类的血吸虫种类，包括曼氏血吸虫（*S. mansoni*）、埃及血吸虫（*Schistosoma haematobium*）和日本血吸虫（*Schistosoma japonicum*），它们可以引起慢性感染并导致人体衰弱，如肝、肠损伤（曼氏血吸虫）或尿道损伤（埃及血吸虫和日本血吸虫）。

曼氏血吸虫是 WHO/TDR 第一批资助的寄生虫测序计划的对象之一（血吸虫基因组网络，http://www.nhm.ac.uk/hosted_sites/schisto/）。因为血吸虫基因组很大（270Mb），早期主要产生 EST 以方便发现基因，值得注意的是，早期就有来自血吸虫流行国家科学家参与研究[91]。曼氏血吸虫基因组（270Mb）正由 TIGR 采用全基因组鸟枪法策略进行测序，采用大片段（BAC）、中等片段（12～15kb）和小片段插入文库（2～3kb）。至今，已产生约31 000个 BAC 克隆的末端序列（http://www.tigr.org/tdb/

e2k1/sma1/)，对小克隆的测序已达到2倍覆盖率（Najib El-Sayed，个人通讯)。基因组3倍覆盖率测序的资金已经到位，这将提供90%以上寄生虫基因的部分或全长序列。由血吸虫基因组网络公布的大量EST数据和排列起来的EST，将对基因组序列的注解具有重要价值[92]。

线虫纲

马来布鲁线虫（*Brugia malayi*）和班氏吴策线虫（*Wucheria bancrofti*）引起淋巴腺丝虫病，使人外貌损伤和衰弱，感染人数超过1亿。另一种线虫，旋盘尾丝虫（*Onchocerca volvulus*)，在亚撒哈拉非洲的一个狭长区域内引起河盲症（river blindness)。由WHO/TDR资助的丝虫基因组计划（http://nema.cap.ed.ac.uk/fgn /filgen1.html;)[93]对这些寄生虫的基因组测序，尤其是EST序列的测定，大多研究是马来布鲁线虫，因为它生活史的各个阶段，包括在蚊寄主中，都可以在实验室进行培养，不过也获得了旋盘尾丝虫的一些数据[94]。

对马来布鲁线虫110Mb基因组的测序已经启动(http: www.tigr.org/tdb/e2k1/bma1/)，测序策略采用了小片段、中等长度片段和大片段（BAC）插入文库，已经获得了1倍覆盖率的序列，桑格研究所在对基因组中的150kb区域进行测序(http://www.sanger.ac.uk /Projects/B _ malayi/)。

结论

从最简单的原核生物到人类，基因组测序提供了各种生物体的大量信息。基因组序列已经对寄生虫学产生了深远影响，从而加深了对寄生虫生物学和寄主-寄生虫相互作用的理解，这些理解是难以通过其他方法得到，基因组序列为实验室研究提供了成千上万个新研究起点。

然而，对蜗居于实验室的科学家，浏览、查询基因组序列并把这些信息应用到手头上的实验中还不是那么容易。为了解决这些问题，已经开发了大量数据库和软件，以方便基因组序列信息的存储和操作，而大部分数据库和软件都是在因特网上公开的。生物特异性数据库（如果蝇库 FlyBase，http://flybase.bio.indiana.edu/[95]和酵母基因组数据库,http://www.yeastgenome.org/）是模式生物的信息总汇，它们不仅提供基因组序列信息，而且还提供株系、突变体、试剂等各方面的信息。寄生虫方面也有这样的几个网站，在科学团体中非常受欢迎并十分有用，例如，PlasmoDB（www.plasmodb.org)，包括所有疟原虫基因组测序项目中原始和最终基因组序列和注解。

对寄生虫基因组测序项目已有大量投资，为了收取回报并保证序列信息能有效地用于基础研究和控制寄生虫疾病，继续发展类似的所有寄生虫数据资源是非常有必要的。还须建立规章制度以保证数据库中的信息能持续更新，反映新的实验发现，并不断改进用于基因搜寻和注解的生物信息学方法。

最后，生物信息学培训，如WHO/TDR疟疾研究和文献试剂资源中心（http: //www.malaria.mr4.org/index.html）提供的培训，对于发达国家和发展中国家的科学

家，获取和利用基因组信息进行寄生虫学基础和应用研究是必要的。

致谢

我对在 TIGR 的同事们表示感谢，尤其要感谢寄生虫基因组成员（Vish Nene, Brendan Loftus, Jane Carlton, Najib El-Sayed, Elodie Ghedin, Ruobing Wang）的支持和鼓励。TIGR 寄生虫基因组测序由国家过敏和感染性疾病研究所（National Institute for Allergy and Infectious Diseases）、The Burroughs Wellcome 基金、美国国防部、国际家畜研究所和 J. Craig Venter 资助。

（欧阳立明 译）

参考文献

1. Sachs J, Malaney P. The economic and social burden of malaria. Nature 2002; 415:680–685.
2. Al-Olayan EM, Beetsma AL, Butcher GA, Sinden RE, Hurd H. Complete development of mosquito phases of the malaria parasite in vitro. Science 2002; 295:677–679.
3. Kocken CH, Ozwara H, van der Wel A, Beetsma AL, Mwenda JM, Thomas AW. *Plasmodium knowlesi* provides a rapid in vitro and in vivo transfection system that enables double-crossover gene knockout studies. Infect Immun 2002; 70:655–660.
4. Johnston DA, Blaxter ML, Degrave WM, Foster J, Ivens AC, Melville SE. Genomics and the biology of parasites. Bioessays 1999; 21:131–147.
5. Singh U, Brewer JL, Boothroyd JC. Genetic analysis of tachyzoite to bradyzoite differentiation mutants in *Toxoplasma gondii* reveals a hierarchy of gene induction. Mol Microbiol 2002; 44: 721–733.
6. Ben Mamoun C, Gluzman IY, Hott C, et al. Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. Mol Microbiol 2001; 39:26–36.
7. Hayward RE, Derisi JL, Alfadhli S, Kaslow DC, Brown PO, Rathod PK. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. Mol Microbiol 2000; 35:6–14.
8. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol 2003; 4:R9.
9. Blader IJ, Manger ID, Boothroyd JC. Microarray analysis reveals previously unknown changes in *Toxoplasma gondii*-infected human cells. J Biol Chem 2001; 276:24,223–24,231.
10. Cleary MD, Singh U, Blader IJ, Brewer JL, Boothroyd JC. *Toxoplasma gondii* asexual development: identification of developmentally regulated genes and distinct patterns of gene expression. Eukaryot Cell 2002; 1:329–340.
11. Florens L, Washburn MP, Raine JD, et al. A proteomic view of the *Plasmodium falciparum* life cycle. Nature 2002; 419:520–526.
12. Lasonder E, Ishihama Y, Andersen JS, et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature 2002; 419:537–542.
13. Cohen AM, Rumpel K, Coombs GH, Wastling JM. Characterisation of global protein expression

by two-dimensional electrophoresis and mass spectrometry: proteomics of *Toxoplasma gondii*. Int J Parasitol 2002; 32:39–51.

14. Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002; 419:498–511.
15. Gardner MJ, Shallom S, Carlton JM, et al. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. Nature 2002; 419:531–534.
16. Hall N, Pain A, Berriman M, et al. Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. Nature 2002; 419:527–531.
17. Hyman RW, Fung E, Conway et al. Sequence of *Plasmodium falciparum* chromosome 12. Nature 2002; 419:534–537.
18. Gardner MJ, Tettelin H, Carucci DJ, et al. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 1998; 282:1126–1132.
19. Bowman S, Lawson D, Basham D, et al. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature 1999; 400:532–538.
20. Hoffman SL, Bancroft WH, Gottlieb M, et al. Funding for malaria genome sequencing. Nature 1997; 387:647.
21. Walliker D, Quayki I, Wellems TE, McCutchan TF. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science 1987; 236:1661–1666.
22. Jing J, Aston C, Zhongwu L, et al. Optical mapping of *Plasmodium falciparum* chromosome 2. Genome Res 1999; 9:175–181.
23. Lai Z, Jing J, Aston C, et al. A shotgun optical map of the entire *Plasmodium falciparum* genome. Nature Genetics 1999; 23:309–313.
24. Salzberg SL, Pertea M, Delcher A, Gardner MJ, Tettelin H. Interpolated Markov models for eukaryotic gene finding. Genomics 1999; 59:24–31.
25. Cawley SE, Wirth AI, Speed TP. Phat—a gene finding program for *Plasmodium falciparum*. Mol Biochem Parasitol 2001; 118:167–174.
26. Jomaa H, Wiesner J, Sanderbrand S, et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 1999; 285:1573–1576.
27. Surolia N, Surolia A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. Nat Med 2001; 7:167–173.
28. Surolia N, RamachandraRao SP, Surolia A. Paradigm shifts in malaria parasite biochemistry and anti-malarial chemotherapy. Bioessays 2002; 24:192–196.
29. Macilwain C. Biologists challenge sequencers on parasite genome publication. Nature 2000; 405: 601–602.
30. Gottlieb M, McGovern V, Goodwin P, Hoffman S, Oduola A. Please don't downgrade the sequencers' role. Nature 2000; 406:121–122.
31. Kissinger JC, Brunk BP, Crabtree J, et al. The *Plasmodium* genome database. Nature 2002; 419: 490–492.
32. Bahl A, Brunk B, Crabtree J, et al. PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. Nucleic Acids Res 2003; 31:212–215.
33. Chakrabarti D, Reddy GR, Dame JB, et al. Analysis of expressed sequence tags from *Plasmodium falciparum*. Mol Biochem Parasitol 1994; 66:97–104.
34. Patankar S, Munasinghe A, Shoaibi A, Cummings LM, Wirth DF. Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. Mol Biol Cell 2001; 12:3114–3125.
35. Munasinghe A, Patankar S, Cook BP, et al. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. Mol Biochem Parasitol 2001; 113:23–34.

36. Mendis K, Sina BJ, Marchesini P, Carter R. The neglected burden of *Plasmodium vivax* malaria. Am J Trop Med Hyg 2001; 64:97–106.
37. Baird JK, Basri H, Purnomo, et al. Resistance to chloroquine by *Plasmodium vivax* in Irian Jaya, Indonesia. Am J Trop Med Hyg 1991; 44:547–552.
38. Schuurkamp GJ, Spicer PE, Kereu RK, Bulungol PK, Rieckmann KH. Chloroquine-resistant *Plasmodium vivax* in Papua New Guinea. Trans R Soc Trop Med Hyg 1992; 86:121–122.
39. Phillips EJ, Keystone JS, Kain KC. Failure of combined chloroquine and high-dose primaquine therapy for *Plasmodium vivax* malaria acquired in Guyana, South America. Clin Infect Dis 1996; 23:1171–1173.
40. Horuk R, Chitnis CE, Darbonne WC, et al. A receptor for the malarial parasite *Plasmodium vivax*: the erythrocyte chemokine receptor. Science 1993; 261:1182–1184.
41. Fraser T, Michon P, Barnwell JW, et al. Expression and serologic activity of a soluble recombinant *Plasmodium vivax* Duffy binding protein. Infect Immun 1997; 65:2772–2777.
42. Trager W, Jensen W. Cultivation of malaria parasites. Nature 1978; 273:621–622.
43. Haynes JD, Diggs CL, Hines FA, Desjardins RE. Culture of human malaria parasites *Plasmodium falciparum*. Nature 1976; 263:767–769.
44. Carlton JM-R, Galinski MR, Barnwell JW, Dame JB. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. Mol Biochem Parasitol 1999; 101:23–32.
45. Carlton JM, Angiuoli SV, Suh BB, et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature 2002; 419:512–519.
46. Carlton JM, Muller R, Yowell CA, et al. Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. Mol Biochem Parasitol 2001; 118:201–210.
47. del Portillo HA, Fernandez-Becerra C, Bowman S, et al. A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. Nature 2001; 410:839–842.
48. Janssen CS, Barrett MP, Lawson D, et al. Gene discovery in *Plasmodium chabaudi* by genome survey sequencing. Mol Biochem Parasitol 2001; 113:251–260.
49. Janse CJ, Carlton JM-R, Walliker D, Waters AP. Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. Mol Biochem Parasitol 1994; 68:285–296.
50. Carlton JMR, Vinkenoog R, Waters AP, Walliker D. Gene synteny in species of *Plasmodium*. Mol Biochem Parasitol 1998; 93:285–294.
51. Tchavtchitch M, Fischer K, Huestis R, Saul A. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. Mol Biochem Parasitol 2001; 118:211–222.
52. Collins WE, Contacos PG, Krotoski WA, Howard WA. Transmission of four Central American strains of *Plasmodium vivax* from monkey to man. J Parasitol 1972; 58:332–335.
53. Carlton J. The *Plasmodium vivax* genome sequencing project. Trends Parasitol 2003; 19:227–231.
54. van Lin LH, Janse CJ, Waters AP. The conserved genome organisation of non-falciparum malaria species: the need to know more. Int J Parasitol 2000; 30:357–370.
55. van Lin LH, Pace T, Janse CJ, et al. Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. Nucleic Acids Res 2001; 29:2059–2068.
56. Rogers WO, Weiss WR, Kumar A, et al. Protection of rhesus macaques against lethal *Plasmodium knowlesi* malaria by a heterologous DNA priming and poxvirus boosting immunization regimen. Infect Immun 2002; 70:4329–4335.
57. Jones SH, Lew AE, Jorgensen WK, Barker SC. *Babesia bovis*: genome size, number of chromosomes and telomeric probe hybridisation. Int J Parasitol 1997; 27:1569–1573.
58. Norval RAI, Perry BD, Young AS. The Epidemiology of Theileriosis in Africa. New York:

Academic, 1992.

59. Mukhebi A, Perry BD, Kruska R. Estimated economics of theileriosis in Africa. Prevent Vet Med 1992; 12:73–85.
60. Dobbelaere D, Heussler V. Transformation of leukocytes by *Theileria parva* and *T. annulata*. Annu Rev Microbiol 1999; 53:1–42.
61. Roos DS, Darling J, Reynolds MG, Hager KM, Striepen B, Kissinger JC. *Toxoplasma* as a model parasite: apicomplexan biochemistry, cell biology, molecular genetics, genomics and beyond. In: Tschudi C, Pearce EJ (eds). Biology of Parasitism. Boston: Kluwer, 2000, pp. 143–167.
62. Sibley LD, Boothroyd JC. Construction of a molecular karyotype for *Toxoplasma gondii*. Mol Biochem Parasitol 1992; 51:291–300.
63. Kohler S, Delwiche CF, Denny PW, et al. A plastid of probable green algal origin in apicomplexan parasites. Science 1997; 275:1485–1489.
64. Manger ID, Hehl A, Parmley S, et al. Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: identification of developmentally regulated genes. Infect Immun 1998; 66:1632–1637.
65. Ajioka JW. *Toxoplasma gondii*: ESTs and gene discovery. Int J Parasitol 1998; 28:1025–1031.
66. Wan KL, Blackwell JM, Ajioka JW. *Toxoplasma gondii* expressed sequence tags: insight into tachyzoite gene expression. Mol Biochem Parasitol 1996; 75:179–186.
67. Ajioka JW, Boothroyd JC, Brunk BP, et al. Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. Genome Res 1998; 8:18–28.
68. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. ToxoDB: accessing the *Toxoplasma gondii* genome. Nucleic Acids Res 2003; 31:234–236.
69. Spano F, Crisanti A. *Cryptosporidium parvum*: the many secrets of a small genome. Int J Parasitol 2000; 30:553–565.
70. Caccio S, Camilli R, La Rosa G, Pozio E. Establishing the *Cryptosporidium parvum* karyotype by *Not*I and *Sfi*I restriction analysis and Southern hybridization. Gene 1998; 219:73–79.
71. Blunt DS, Khramtsov NV, Upton SJ, Montelone BA. Molecular karyotype analysis of *Cryptosporidium parvum*: evidence for eight chromosomes and a low-molecular-size molecule. Clin Diagn Lab Immunol 1997; 4:11–13.
72. Strong WB, Nelson RG. Preliminary profile of the *Cryptosporidium parvum* genome: an expressed sequence tag and genome survey sequence analysis. Mol Biochem Parasitol 2000; 107:1–32.
73. Liu C, Vigdorovich V, Kapur V, Abrahamsen MS. A random survey of the *Cryptosporidium parvum* genome. Infect Immun 1999; 67:3960–3969.
74. Abrahamsen MS. *Cryptosporidium parvum* gene discovery. Adv Exp Med Biol 1999; 473: 241–247.
75. Zhu G, Marchewka MJ, Keithly JS. *Cryptosporidium parvum* appears to lack a plastid genome. Microbiology 2000; 146(Pt 2):315–321.
76. Foth BJ, Ralph SA, Tonkin CJ, et al. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. Science 2003; 299:705–708.
77. Ivens AC, Lewis SM, Bagherzadeh A, Zhang L, Chan HM, Smith DF. A physical map of the *Leishmania major* Friedlin genome. Genome Res 1998; 8:135–145.
78. Myler PJ, Audleman L, deVos T, et al. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci USA 1999; 96:2902–2906.
79. Dubessay P, Ravel C, Bastien P, et al. The switch region on *Leishmania major* chromosome 1 is not required for mitotic stability or gene expression, but appears to be essential. Nucleic Acids Res 2002; 30:3692–3697.
80. El-Sayed NM, Ghedin E, Song J, et al. The sequence and analysis of *Trypanosoma brucei* chromosome II. Nucleic Acids Res 2003; 31:4856–4863.

81. Hall N, Berriman M, Lennard NJ, et al. The DNA sequence of chromosome 1 of an African trypanosome: gene content, chromosome organisation, recombination and polymophism. Nucleic Acids Res 2003; 31:4864–4873.
82. Gull K. The biology of kinetoplastid parasites: insights and challenges from genomics and post-genomics. Int J Parasitol 2001; 31:443–452.
83. Zingales B, Rondinelli E, Degrave W, et al. The *Trypanosoma cruzi* genome initiative. Parasitol Today 1997; 13:16–22.
84. Brandao A, Urmenyi T, Rondinelli E, Gonzalez A, de Miranda AB, Degrave W. Identification of transcribed sequences (ESTs) in the *Trypanosoma cruzi* genome project. Mem Inst Oswaldo Cruz 1997; 92:863–866.
85. Aguero F, Verdun RE, Frasch AC, Sanchez DO. A random sequencing approach for the analysis of the *Trypanosoma cruzi* genome: general structure, large gene and repetitive DNA families, and gene discovery. Genome Res 2000; 10:1996–2005.
86. Katinka MD, Duprat S, Cornillot E, et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 2001; 414:450–453.
87. Yu LZ, Birky CW Jr, Adam RD. The two nuclei of *Giardia* each have complete copies of the genome and are partitioned equationally at cytokinesis. Eukaryot Cell 2002; 1:191–199.
88. McArthur AG, Morrison HG, Nixon JE, et al. The *Giardia* genome project database. FEMS Microbiol Lett 2000; 189:271–273.
89. Bhattacharya A, Satish S, Bagchi A, Bhattacharya S. The genome of *Entamoeba histolytica*. Int J Parasitol 2000; 30:401–410.
90. Mann BJ. *Entamoeba histolytica* Genome Project: an update. Trends Parasitol 2002; 18:147–148.
91. Franco GR, Valadao AF, Azevedo V, Rabelo EM. The *Schistosoma* gene discovery program: state of the art. Int J Parasitol 2000; 30:453–463.
92. Oliveira G, Johnston DA. Mining the schistosome DNA sequence database. Trends Parasitol 2001; 17:501–503.
93. Williams SA, Lizotte-Waniewski MR, Foster J, et al. The filarial genome project: analysis of the nuclear, mitochondrial and endosymbiont genomes of *Brugia malayi*. Int J Parasitol 2000; 30:411–419.
94. Williams SA, Laney SJ, Lizotte-Waniewski M, Bierwert LA, Unnasch TR. The river blindness genome project. Trends Parasitol 2002; 18:86–90.
95. FlyBase Consortium. The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 2003; 31:172–175.

21

极端嗜盐古生菌基因组分析

Shiladitya DasSarma

引言

极端嗜盐生物是一类新颖的微生物，其最适生长要求的盐度是海水的5～10倍（即3～5mol/L NaCl）[1,2]，包括多种原核生物——古生菌和细菌，以及某些真核生物。极端嗜盐生物发现于近海的高盐环境，或近海的或非海洋的盐沉积环境。两个最大的高盐湖孵育了各种各样的嗜盐生物，一个是美国西部的大盐湖（Great Salt Lake），另一个是中东的死海，最有趣的高盐环境是用海水晒盐的小盐场，它们分布于世界各地。一些高盐环境随着时间的推移呈现出逐渐增加的盐度梯度，并使得越来越嗜盐的种类依次生长，包括复杂的微生物垫、呈亮红或橘红色的壮观的水华。这些环境具有重要的生态意义，它们常常抚养着像粉红火烈鸟这样的奇特生物，这些鸟的颜色来自带颜色的嗜盐微生物。嗜盐微生物的一个典型特征，是它们细胞内有高浓度相溶性溶质（例如氨基酸、多羟基聚合物和盐等）作为渗透压保护剂（osmoprotectant），因此，在高盐环境中细胞不被裂解。

尽管已能培养多种嗜盐微生物，但至今仅有一株极端嗜盐菌——盐杆菌（*Halobacterium*）菌株NRC-1完成了基因组测序[3,4]，该菌是大盐湖和晒盐场等多种高盐环境中的常见菌。种系进化分析表明，它属生物三域之一：古生菌（图1），最适生长盐度为4.5mol/L NaCl，接近饱和点，胞内还含高浓度K^+。菌株NRC-1是一株嗜中温古生菌，最适生长温度42℃。尽管盐杆菌属的生理能力有限，但菌株NCR-1的代谢富于多样性，可以有氧、厌氧或靠光合作用生长。光养生长是通过细菌视紫红质（bacteriorhodopsin）的光驱动质子泵介导，它在紫质膜中形成两维晶格结构，菌株NRC-1有很强的抗紫外线和γ射线的能力，并表现出复杂的运动性反应，包括趋光、趋化、气泡介导的漂浮。菌株NRC-1的最重要特征是通过基因组分析发现它有一个高度酸性蛋白质组，这对蛋白质维持在高盐环境中的溶解性和功能是必不可少的。有意义的是，该菌适于用现代遗传学方法，如基因敲除、表达载体和互补系统等来研究分析，这使得该菌株成为研究极端生物和古生菌功能基因组的良好模式生物[2]。

除了菌株NRC-1外，正在进行基因组分析的还有其他几种嗜盐生物，其中最值得注意的是两种死海古生菌，死海盐盒菌（*Haloarcula marismortui*）和沃氏富盐菌（*Haloferax volcanii*）[1]，它们比菌株NRC-1的嗜盐性稍低，最适盐度为2～3mol/L NaCl，并对镁离子有很高抗性，这反应了它们生活环境的盐分组成，它们在以简单的糖和碳水化合物作为碳源和能源的培养基中也表现出代谢能力。其他几个值得研究的嗜盐生物类别，包括嗜碱嗜盐生物，它们在pH9.0～11.0的碱水湖中生长、生活，在南

极湖中结冰温度下的嗜冷嗜盐生物、耐受广盐度的嗜盐细菌；还有真核嗜盐生物，如绿藻（*Dunaliella salina*）。最后一提的是，一个几乎与 NRC-1 染色体完全一样的嗜盐古生菌的测序正在进行中。当前有关嗜盐生物基因组计划名单可从马里兰大学生物技术所海洋生物技术中心网页上了解到：Halophile Genomes Web(http://zdna2. umbi. umd. edu)。

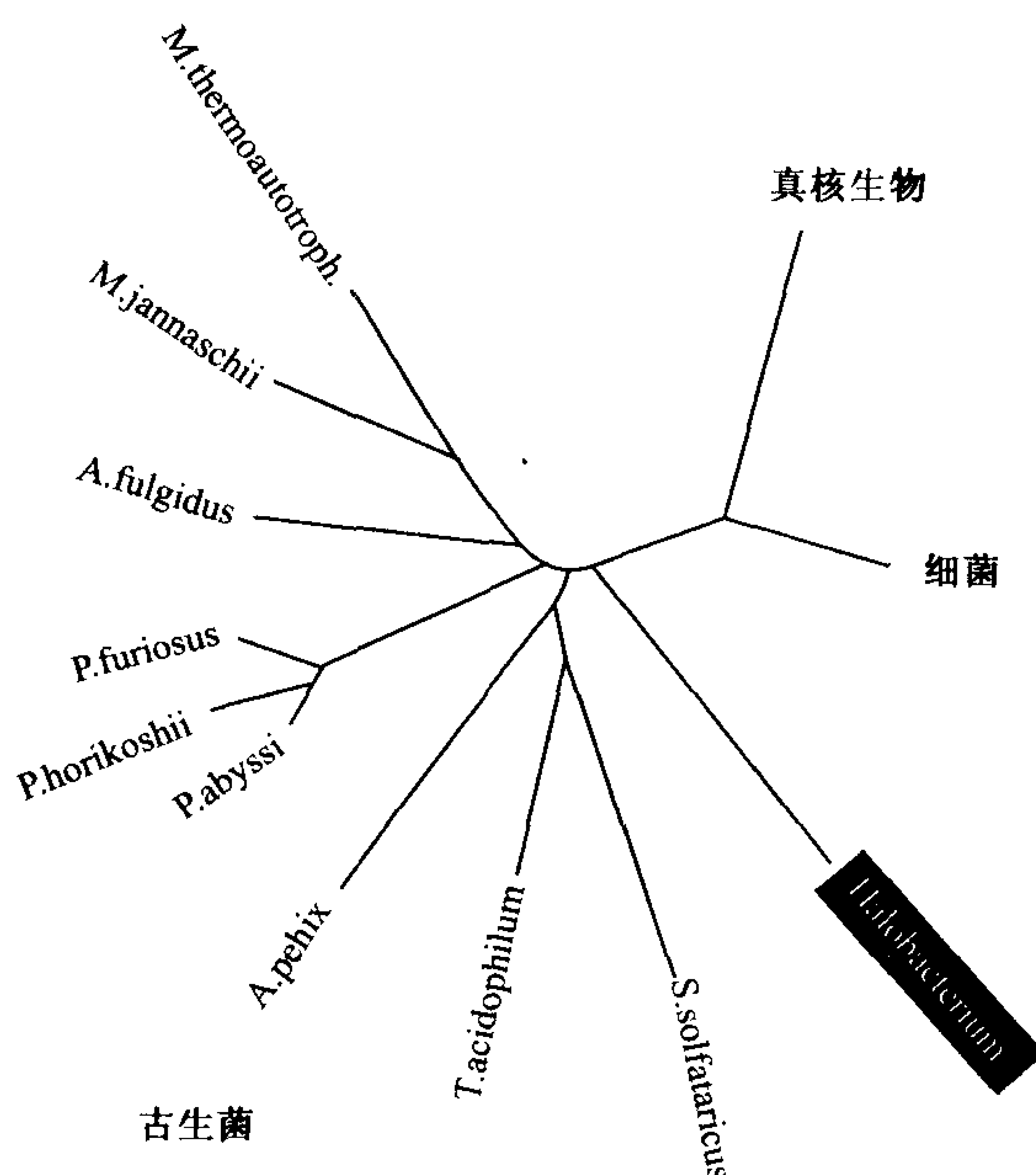


图 1　几种古生菌的全基因组进化树。用 SHOT 网络服务器通过邻接法（neighbor-jioning）分析了所含基因的种系发生，表明盐杆菌定位于种系进化树古生菌分支的基部。

盐杆菌基因组

半世纪前开始了盐杆菌的基因组研究，其基因组分两部分，即富含 G+C 主组分和 A+C 含量相对多的（58%G+C）卫星 DNA[5]。随后的研究表明，这些卫星 DNA 主要是一些大染色体外复制子，它们含有许多可转座的插入序列（IS）[6]。对盐杆菌（*Halobaterium*）NRC-1 的大量作图，揭示有三种复制子：pNRC100 约 200kbp、pNRC200 为 pNRC100 的近两倍和一个 2Mb 染色体（图 2）[7,8]。发现 pNRC100 复制子与 pNRC200 部分同源且以倒位体形式存在[7]。对菌株 NRC-1 染色体和另一株野生菌 GRB 进行了限制性图谱比较，显示大量区域有同源性，少数区域有差异，包括一个大倒位序列和一个插入序列。对整理好次序的覆盖了盐杆菌 *GRB* 和沃氏富盐菌（*H. volcanii*）基因组的黏粒文库进行了杂交比较，没有发现任何保守基因组合[9]，这些及其他的基因图谱计划都表明，盐古生菌基因组间存在明显的多样性。

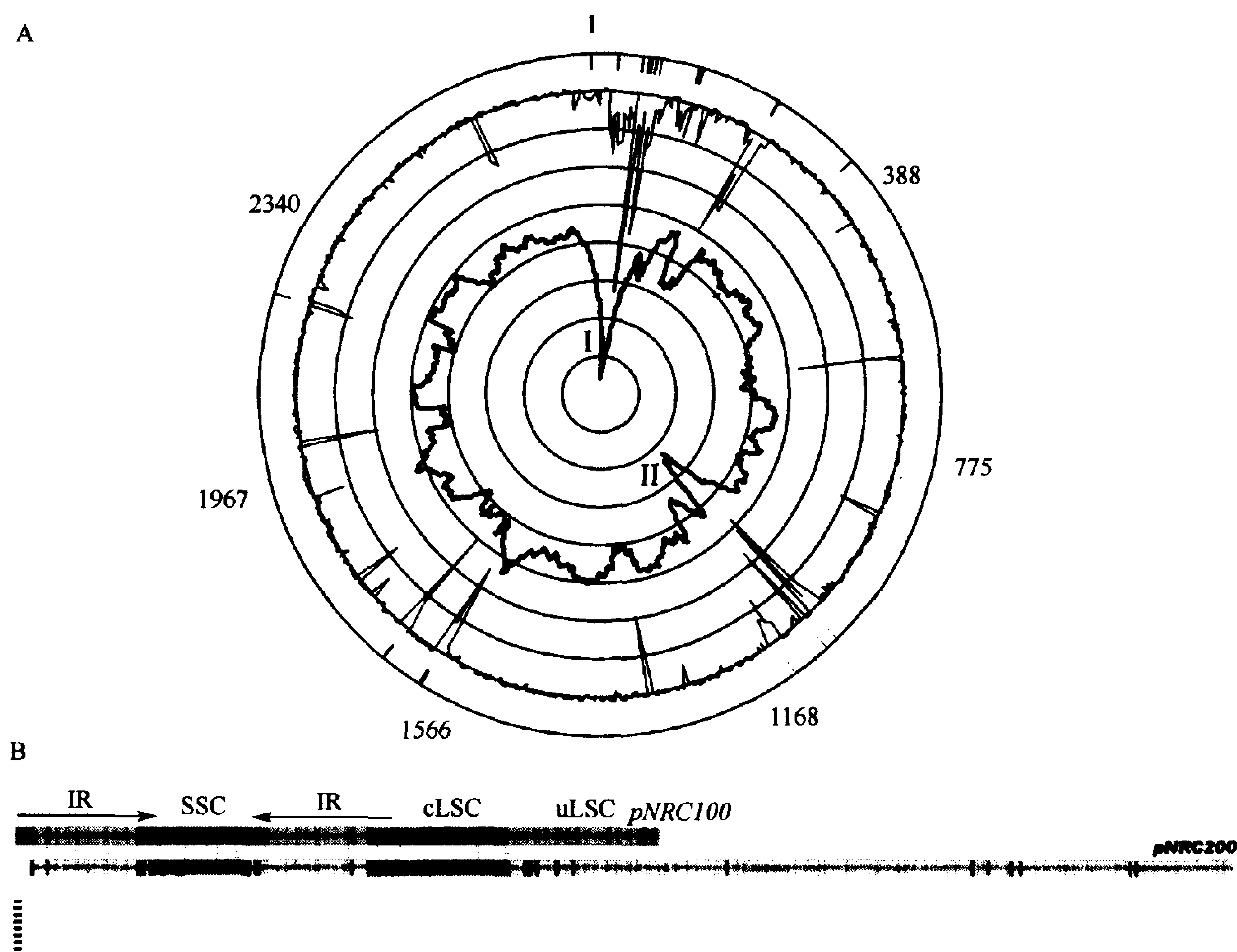


图 2　A. 盐杆菌（*Halobaterium*）NRC-1 大染色体环状图谱。B. pNRC100 和 pNRC200 复制子的线状遗传图谱比对。A. 大染色体环状图包含 IS 元件的位置（外圈），χ 平方分析（红线），可读框 G+C 组成（黑线）。与最外圈相连彩色线段表示染色体 IS 位置（ISH1，浅褐色；ISH2，紫色；ISH3，绿色；ISH4，黄色；ISH6，粉色；ISH8，蓝色；ISH10，红色）。罗马数字 I、II 表示富 AT 岛。B. 以线状形式描述环状复制子，基因和 IS 元件用块状表示。两个复制子包含 145 428 bp 一致区和 45 918 bp（pNRC100）或 219 997 bp（pNRC200）的特有 DNA[3,4]。33kb 到 39kb 的反向重复用黄色（在所有拷贝中保守）和橙色（在一些拷贝中保守）表示；小单拷贝区用紫色表示；共同单个大拷贝区用亮绿色表示；独一无二的单个大拷贝区用棕褐色（pNRC100）和淡绿色（pNRC200）表示。IS 元件用暗橙色（ISH2）、棕色（ISH3）、靛蓝色（ISH5）、蓝色（ISH7）、暗绿色（ISH8）、鬼蓝（ISH9）、蓝灰色（ISH11）。两个 pNRC 复制子含 69 个 IS 元件（44 个是独特的），其中 29 个在 pNRC100 上、40 个在 pNRC200 上；有 6 个因子是反向重复（在 pNRC100 和 pNRC200 中都重复 2 次），在 pNRC100 和 pNRC200 中的 SSC 区中都有 4 个因子；在 pNRC100 和 pNRC200 中的共同单个大拷贝区都含 7 个因子；在独特单个大拷贝区含 23 个因子，其中 6 个在 pNRC100 上，17 个在 pNRC200 上。（图 2A 经冷泉港实验室出版社许可复制，文献[11]）（另见文前彩色插图 21-2）。

基因组测序与分析

由于 G+C 含量高和 IS 数量大，NRC-1 基因组分两步测序，首先，pNRC-100 复制子通过对纯化共价闭环 DNA 文库的鸟枪法随机测序，并对已克隆和构建物理图谱的 *Hind* III 片段进行定向测序相结合完成[3, 7]。通过这种策略把经常发生倒位等基因重排，并含有很多 IS 元件的不稳定复制子完整地组装起来。随后，对整个基因组进行鸟

枪随机测序，涵盖相对稳定大染色体 DNA 的 7.5 倍。其余不好测的部分通过 PCR 和引物步移（primer walking）的方法获得序列。用 *Phred*、*Phrap* 和 *Consed* 软件对 NRC-1 基因组实行组装，开始先屏蔽所有已知和推测的新 IS 元件，以避免嵌合重叠群的形成[4, 10]。

菌株 NRC-1 的全部基因组序列为 2 571 010bp，包括 2 014 239bp 富含 G+C 的染色体和两个小环状 DNA，191 346bp 的 pNRC100 与 365 425bp 的 pNRC200（表 1；图 2）[3,4]。有趣的是，pNRC100 与 pNRC200 有一段长度为 145 428bp 的 100%同源区，包括 33～39kb 的反向重复，该重复序列介导了倒位异构化；还包括小单拷贝区和大单拷贝区的一部分（图 2）[7]。这个独特大单拷贝区在 pNRC100 中为 45 918bp，而在 pNRC200 中为 219 997bp。用 Glimmer（Gene Locator and Interpolated Markov Modeler）发现基因组中可能有 2630 个基因，其中 64%编码蛋白与数据库中明显同源[4]，此外，还发现 52 种 RNA 基因。在 pNRC100 和 pNRC200 中，有约 40 种基因编码的蛋白对细胞生存是必要或重要的，如一种 DNA 聚合酶，TBP 和 TFB 转录因子以及精氨酰 tRNA 合成酶，表明这两种复制子应属于小染色体而非大质粒[3,4]。

表 1　盐杆菌（*Halobaterium*）NRC-1 基因组统计分析

项目	总共	染色体	pNRC200	pNRC100
大小（bp）	2 571 010	2 014 239	365 425	191 346
G+C（%）	65.9	67.9	59.2	57.9
预测基因数	2682	2111	374	197
编码（%）	84	87	76	71
IS 因子数量	91	22	40	29
ISH1	1	1	0	0
ISH2	13	4	5	4
ISH3	23	5	10	8
ISH4	2	1	0	1
ISH5	6	0	4	2
ISH6	2	1	1	0
ISH7	4	0	2	2
ISH8	21	5	10	6
ISH9	4	0	2	2
ISH10	6	2	2	2
ISH11	7	2	3	2
ISH12	2	1	1	0

蛋白质组分析

盐杆菌（*Halobaterium*）NRC-1 基因组序列分析中，最引人注目的结果是，它编码的蛋白质均为极端酸性，这直接与在高盐（>4mol/L KCl）胞质中行使蛋白功能有关[11]。预测蛋白的平均等电点值（p*I*s）约为 5，并以后被蛋白组分析所证实（图 3）。类似地，从死海盐盒菌（*H. marismortui*）和沃氏富盐菌（*H. volcanii*）两嗜盐生物的部分基因组预测，其蛋白质组也是酸性，相反，几乎所有其他蛋白质组的平均 p*I*s 都接近中性。有几个例外值得注意，其一，热自养甲烷杆菌（*Methanobacterium thermoautotrophicum*）含有一个酸性蛋白质组和一个相对高的内部 K^+ 浓度（~1mol/L）；还有三种超嗜热生物需氧热棒菌（*Pyrobaculum aerophilum*）、激烈热球菌（*Pyrococcus furiosus*）和硫磺矿硫化叶菌（*Sulfolobus solfataricus*），它们有相对碱性的蛋白质组。同源性建模表明菌株 *NRC*-1 蛋白的酸性 p*I* 与其表面高浓度负电荷有关[11]，例如，转录因子 TbpE 与拓扑异构酶亚基 GyrA 与非嗜盐生物中的同源蛋白相比，其表面负电荷有显著增加[11]。

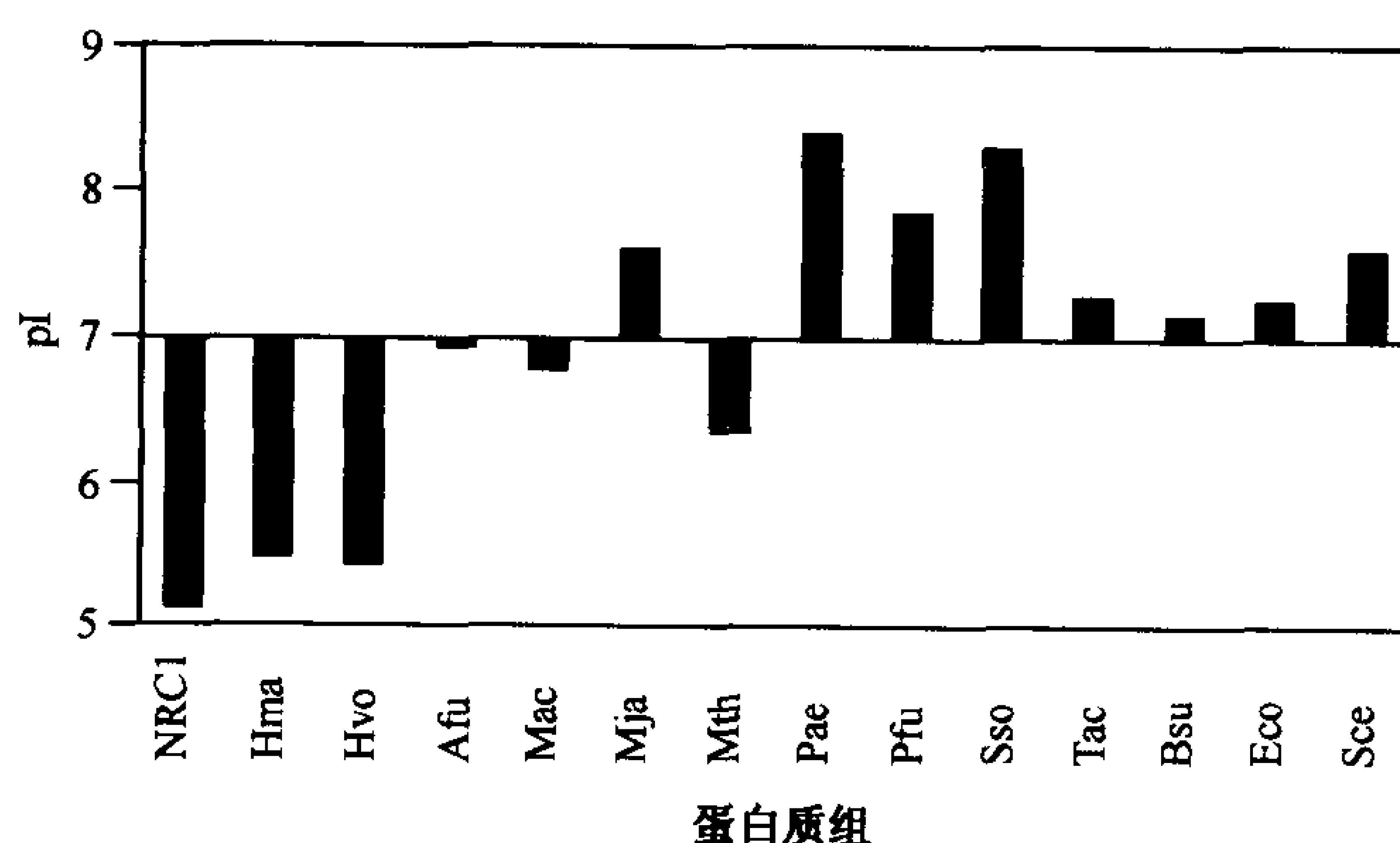


图 3 由基因组序列预测的蛋白质组平均 p*I* 图谱：盐杆菌（*Halobacterium* sp）NRC-1（NRC-1）、死海盐盒菌［*Haloarcula marismortui*（Hma）］、沃氏富盐菌［*Haloferax volcanii*（Hvo）］，闪烁古生球菌［*Archaeoglobus fulgidus*（Afu）］、噬乙酸甲烷八叠球菌［*Methanosarcina acetivorans*（Mac）］、詹氏甲烷球菌［*Methanococcus jannaschii*（Mja）］、热自养甲烷杆菌［*Methanobacterium thermoautotrophicum*（Mth）］、需氧热棒菌［*Pyrobaculum aerophilum*（Pae）］、激烈火球菌［*Pyrococcus fuiosus*（Pfu）］、硫磺矿硫化叶菌［*Sulfolobus solfataricus*（Sso）］、嗜酸热原体［*Thermoplasma acidophilum*（Tac）］、枯草芽孢杆菌［*Bacillus subtilis*（Bsu）］、大肠杆菌［*Escherichia coli* K12（Eco）］、酿酒酵母［*Saccharomyces cerevisiae*（Sce）］。

G+C 组成与 IS 元件

嗜盐生物基因组的共同特点是，主要部分为高，G+C 组成，卫星部分为低 G+C，并有大量 IS 元件[6]。菌株 NRC-1 占基因组 22%，两个 pNRC 复制子的 G+C 含量（58%~59%）明显低于大染色体 G+C 含量（68%），并含有基因组中绝大多数

(69%/71%或76%) IS元件(图2)。此外,染色体有两个区的G+C低于平均数,一个区为270kbp (region I),含有65%G+C和13个IS元件;另一个为150kbp区(region II),G+C含量66%、含有4个IS元件(图2)[11]。有趣的是,在pNRC反向重复序列上有一个15kb区,其G+C含量(64%)比整个pNRC100含量(58%)高,而没有任何IS元件[3],说明三种复制子中,均具有不同特征的基因组区域。总之,在NRC-1基因组中,有代表着12类91个IS元件(表1)[4],表明IS元件参与了菌株NRC-1的复制子之间DNA交换。

菌株NRC-1的高G+C含量似乎是对太阳辐射强的一种适应(例如,减少胸腺嘧啶二聚体的形成),统计表明,与一个大小相当、G+C组成为50%的复制子相比,NRC-1大染色体上形成胸腺嘧啶二聚体的位点数要比它低60%。然而,双核酸分析显示,这样的位点比根据G+C含量预测的值还低20%[11],这样高的G+C组成也导致密码子使用中,第三位碱基偏嗜于G+C (86%G+C,而前两位是70%和46%)。

盐杆菌基因组的注释

盐杆菌基因组联盟(The *Halobacterium* Genome Consortium)是一个代表12个机构的国际小组,从1999年夏天到2000年夏天展开了对NRC-1基因组的注释。数据资料从3倍覆盖开始公布到全部完成,并于2000年元月在麻省的Amherst举办了研讨会,这一成果导致对这第一个嗜盐生物序列的彻底分析,并使它最大化地服务于社会。在以后的两年中,大量基因被发现,在此,仅对当前注释的要点作一总结,详细数据可到Halophile Genome网址检索(http://zdna2.umbi.umd.edu)。

DNA复制

从菌株NRC-1基因组发现编码一个古生菌的异二聚体D型DNA聚合酶、很多类似真核的复制蛋白、两种B类DNA聚合酶,其一由pNRC200编码、复制子识别和解旋酶召集蛋白(helicase recruiter) (10 Orc1/Cdc6)、复制解旋酶(MCM)、ssDNA结合蛋白(6Rfa)、引发酶(2 Pri)、滑动钳载体(clamp loader) (RfcABC)、进行性滑动钳(processivity clamp) (2种增殖细胞核抗原类似物)、I型拓扑异构酶(TopA)、II型拓扑异构酶(Top6A和Top6B)、RNA引物切除酶(Rad2和RNaseH)、还有几种参与复制的细菌基因,一种引发酶(DnaG)和拓扑异构酶(GyrA和GyrB)。有趣的是,还发现了编码真核复制起始识别复合体蛋白Orc1/Cdc6基因的多个拷贝,有3个散布于大染色体上,表明存在多复制起始的可能性[11]。

DNA修复

菌株NRC-1基因组包含许多修复基因(图4),对修复由环境中强太阳辐射造成的DNA损伤是必需的[12],NRC-1表现出高水平抗紫外和γ辐射的能力与预测相符。光复活作用(photoreactivation)在盐杆菌中非常有效,基因组编码两种光解酶/蓝光受体(cryptochrome)类似物,其一可能用于DNA修复。碱基的剪切修复是由Ogg、AlkA、MutA和Nth同源物完成,也可能通过XthA(一个IV类AP核酸内切酶的同源物)和

甲基化酶 Ogt（可能参与甲基化损伤修复）来完成。菌株 NRC-1 也编码细菌剪切修复复合体 UvrABCD 同源物。特别有趣的是，在 NRC-1 中出现了编码真核形式的剪切蛋白（Rad2、Rad3、Rad25 与 ERC4）基因，表明它有两套复制修复系统。在菌株 NRC-1 中发现了错配修复蛋白 MutS1、MutS2 和 MutL，还发现了 RadA1、RadA2 和 RecA/Rad51 基因的同源物（可能编码重组酶）、MRE11 以及可能参与同源重组和重组修复的 Holliday 结构解离酶、负责绕过损伤的细菌 UmuC 聚合酶的同源物以及一种真核 ATP 型的 DNA 连接酶。

转录

与其他古生菌类似，NRC-1 中发现了类似于真核 RNA 聚合酶 II 转录系统的简化本，包括 Rpo 亚单位 A、C、B′、B″、E′、E″、H、K、L、N 和 M[4]。此外，惊奇地发现 NRC-1 基因组编码 13 个拷贝的 TBP 和 TFB 转录因子，包括 5 个完整和一个部分的 *tbp* 基因（4 个位于 pNRC100，一个位于 pNRC200，一个位于大染色体上）以及 7 个 *tfb* 基因（两个在 pNRC200 上，5 个在大染色体上）[13]。这些结果表明，可能存在交替利用 TBP-TFB 组合进行启动子选择的基因调节新机制。与该假说相符的是，基因组序列分析 *bop* 和启动子饱和突变证实，存在另外一个 TATA 框[14]，还发现了近 100 种、多数为细菌型的转录调节蛋白。

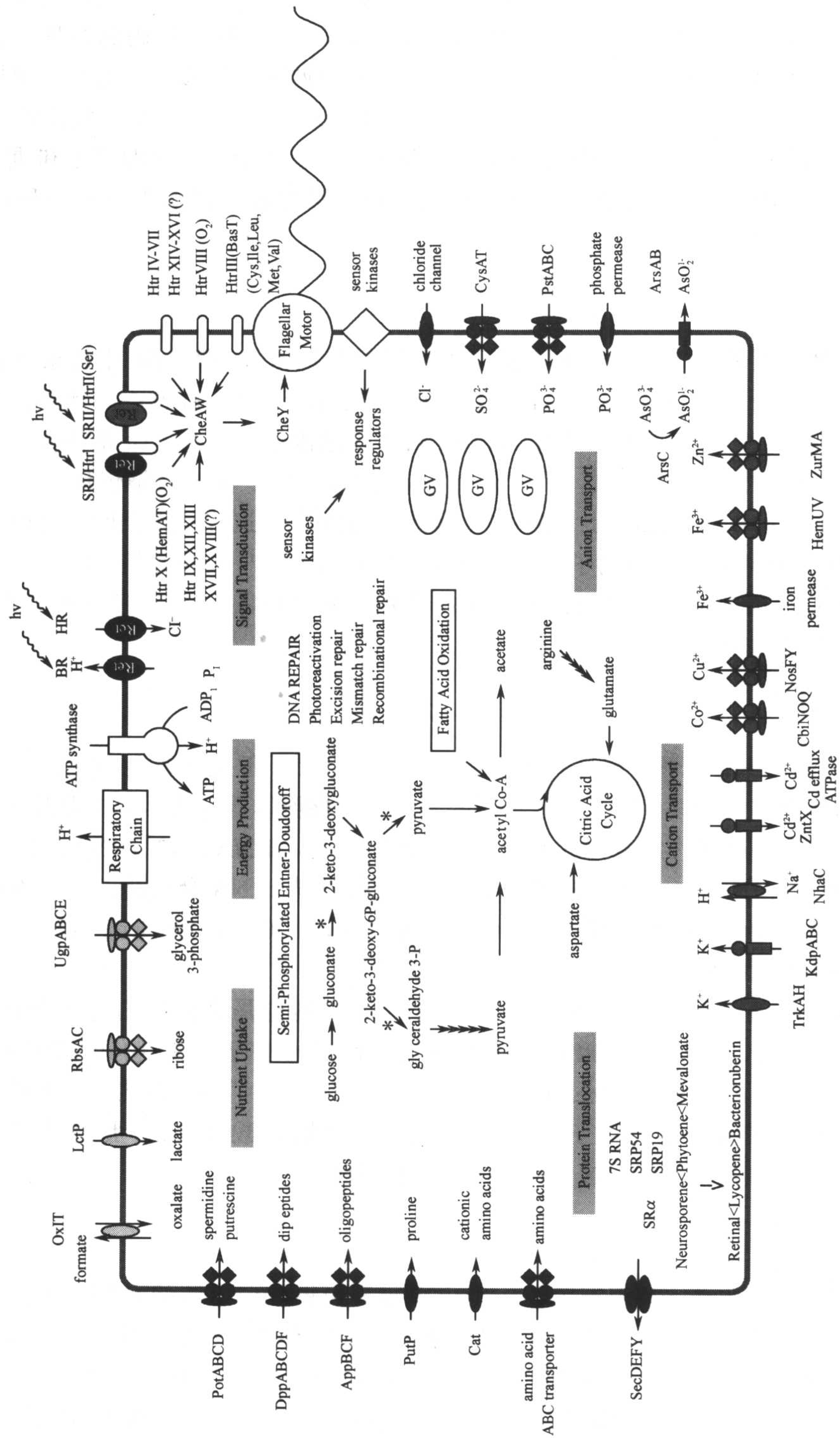
蛋白质合成

菌株 NRC-1 的翻译系统具有真核和细菌的杂合性，但是像其他古生菌一样，它的所有核糖体蛋白与真核生物类似，特别的是，它的核糖体蛋白基因排成类似细菌操纵子的多基因簇。除了 52 种 RNA（16S、23S 和 5S rRNA、47 tRNA、7S RNA 和 RNAaseP）外，NRC-1 基因组还编码 18 种不同的氨酰 tRNA 合成酶，加上谷氨酰胺与天冬酰胺 GatABC 氨基转移酶[4]，其中的一种由 pNRC200 编码的氨酰 tRNA 合成酶 ArgRS，与细菌和酵母线粒体中的酶有紧密亲缘关系。

菌株 NRC-1 蛋白质分泌的通用分泌（Sec）机制也是真核生物与细菌的杂合体，Sec61α、Sc61γ、SRP54、SRP19 和 7SRNA 与真核中相对应的蛋白有亲缘关系，而 FtsY、SecD 和 SecF（但不是 SecA）与细菌蛋白相关[4]。除了 Sec 系统，近来的生物信息学分析表明，细菌中主要用于分泌氧还蛋白的双精氨酸（Tat）蛋白运出途径，在 NRC-1 中也存在，很可能在这个古生菌中被普遍应用[15,16]。

细胞囊

NRC-1 细胞由单一脂双层膜和一个由细胞表面糖蛋白组装的 S 层包围，细胞质能通过与外部环境中 Na^+ 浓度相当的高 K^+ 浓度来维持细胞与外部高盐环境间的渗透压平衡，像其他古生菌一样，基本极性脂是古生菌脂（archeaol），这是一种含有从 C_{20} 类异戊二烯（isopreniod）衍生的植烷（phytanyl）链的甘油二酯脂。菌株 NRC-1 基因组含有类异戊二烯合成的所有关键酶，包括 HMG-辅酶 A 还原酶（MvaA）——生长抑制剂美维诺林（mevinolin）的效应靶点[4]。为了维持离子平衡，NRC-1 编码多种 K^+ 转运蛋白，包括 KdpABC（一种 ATP 驱动的 K^+ 转运系统）以及 TrkAH（一种低亲和性的、



由膜电势驱动的 K^+ 转运蛋白）（图 4）。活跃的 Na^+ 外流可能是由单向 Na^+/H^+ 反向转运蛋白 NhaC 介导的。有趣的是，编码 KdpABC 基因、编码 TrkA 基因的三个拷贝（共五个）与编码 NhaC 基因的一个拷贝（共三个）位于 pNRC200 上。此外，还发现用于营养吸收的主动运输蛋白，有的用于阳离子氨基酸（Cat）和脯氨酸（PutP）、二肽(DppABCDF)、寡肽（AppACF)、糖运转蛋白（Rbs)，有的用于重金属（亚砷酸盐和镉）和其他有毒化合物（多药物抗性类似物）清除，还有磷酸盐运输系统 PstABC 的多个拷贝和磷酸盐透性酶等。

紫质膜

NRC-1 含有紫膜，这是一种具有光驱动质子泵功能的细菌视紫红质组成的二维晶格结构（图 4)；细菌视紫红质是由细菌视蛋白（bacterioopsin）和视黄醛发色团组成的复合体。在强光下，细胞进行光营养生长，这种能力在浮游细菌中是刚被认识到的[12]，发现了簇居于染色体并协同调节的 5 种紫膜调节子基因，包括 *bop*，编码细菌视紫红质；*crt*B1 和 *brp*，分别编码第一步和最后一步视黄醛的合成；*blp*，未知功能基因；以及 *bat*，感受器-调节子[14]。*bat* 基因的产物（Bat）是一个小基因家族的成员，含有一个 GAF 域（结合 cGMP)，PAS/PAC 域（氧化还原感应)，结合 DNA 的 α 螺旋-拐角-α 螺旋模体（motif)，该模体可能结合基因的上游激活序列（upstream activator sequence, UAS）而使基因活化。*bop* 基因的 TATA 框序列起源于古生菌启动子的相同序列，表明在该菌的转录中有新因子参与，如交替的 TBP 和 TFB 蛋白[14]。

趋性和信号传导

盐杆菌及其类似菌有很强的趋化性和趋光性，其游动行为受化学梯度和光密度或色彩的调节。有大量趋性基因被识别，包括编码趋光受体的 *sop*I 和 *sop*II；SRI 和 SRII 属

图 4 盐杆菌（*Halobacterium*）NRC-1 基因组整体图[4]。图中描绘了能量产生、营养吸收、膜组装、阴阳离子转运以及信号传导等几方面的内容。展示了通过由呼吸链的质子运输、由细菌视紫红质(BR，紫色卵形）光驱动的质子泵、由卤紫质（HR，蓝色卵形）进行的氯离子运输等化学渗透耦合(chemiosmotic coupling）产生 ATP 的过程。在下方，展示了半磷酸化 Entner-Doudorff 途径、脂肪酸氧化和三羧酸循环。星号标出的是尚未鉴定的酶。基因组编码了各种各样的营养吸收系统（黄色或棕色标示的结构），如甘油-3-磷酸运转蛋白（UgpABCE）和糖 ABC 运转蛋白（RbsAC)、乳酸盐运转蛋白（LctP)、甲酸盐-草酸盐反向运转蛋白（OxiT)、亚精胺与丁二胺 ABC 运转蛋白(PotABCD)，氨基酸（PutP, Cat）和二肽运转蛋白（DppABCDF)。可能存在以通用 ABC 运转蛋白为代表的其他氨基酸吸收系统。图中黑色所示是蛋白移位器的几个组分（SecDEFY、SRP19、SRP54、SRα)。图中标示出类胡萝卜素和视黄醛（Retinal, Ret）的生物合成。绿色所示是下列几种阳离子运输：K^+（TrkAH 与 KdpABC)、Na^+（NhaC)、Cd^{2+}（ZntX 与 Cd 外排 ATP 酶)、Co^{2+}(CbiNOQ)、Cu^{2+}（NosFY)、Fe^{3+}（铁离子透性酶与 HemUV）和 Zn^{2+}（ZurMA)。红色所示是下列几种阴离子运输：SO_4^{2-}（CysAT)、PO_4^{3-}（PtABC 与磷酸盐透性酶)、Cl^-（氯离子通道）和砷酸盐（ArsABC)。图中标示出由光受体与信号传导组分组成的复杂系统，包括两个感受器（蓝色的是 SRI、橙色的是 SRII)、17 个响应光（hv)、O_2 或氨基酸传导蛋白（HtrI-HtrX、HtrXII-HtrXVIII)。运动性信号经过 CheAW 和 CheY 向鞭毛马达的传递，如箭头所示。鞭毛绘成波浪线。图中列举了一个感受激酶［膜结合的（白色菱形）或胞质的］和响应调节蛋白的例子。细胞内还绘出了气囊(GV，白色卵形）和 DNA 修复系统。(另见文前彩色插图 21-4)

细菌视紫红质家族（并且也包括卤视紫红质-halorhodopsin，一种氯泵）（图 4）[12]。SRI 介导橙黄光的吸引和对近紫外光的排斥响应，而 SRII 是一种蓝光排斥光受体蛋白。有趣的是，嗜盐古生菌的视紫红质类似物在真菌、藻类、海洋细菌和蓝细菌的基因组中都有发现[12]。总共发现 17 种 *htr* 基因，它们编码与细菌趋化受体同源的膜镶嵌蛋白；还发现了一整套编码趋化决定子的 *che* 基因。有 6 种鞭毛蛋白基因和古生菌类型的鞭毛器[16]，一个大基因簇 *fla*D-K 编码着古生菌鞭毛器，其中 *fla*D、*flaE*、*fla*G、*fla*H、*fla*I 和 *fla*J 与其他古生菌相似，只有 *fla*K 与细菌鞭毛调节子相似。显然，在 NRC-1 基因组中存在双组分调节系统（two-component regulatory system），包括 6 个响应调节基因和 14 种组氨酸激酶。NRC-1 基因组还揭示出多个昼夜节律光调节子，包括一个真核蓝光受体和蓝细菌的 KaiC 类似蛋白，这与光营养细菌的昼夜节律一致的[12]。

气囊

盐杆菌属的各个种像许多光营养的水生原核生物一样，具有通过合成充气囊调节浮力的能力（图 4），对气囊形成的条件已在 NRC-1 中通过遗传分析进行了深入研究[17]。在 NRC-1 的两个质粒 pNRC100 和 pNRC200 上，都有一个基因簇 *gvp*MLKJI-HGFEDACN（O）对野生型气囊的合成是必要的。而在对菌株 NRC-1 的基因组序列分析时，发现在 pNRC200 上有沉默、近乎完整的基因簇 *gvp*（只缺少 *gvp*M）[4, 12]。

类胡萝卜素和视黄醛

盐杆菌属能产生红橙色类胡萝卜素，它们对光传导和光损伤保护是必需的，其中最多的是菌红素（bacterioruberin）（图 4）。已发现 NRC-1 中编码细菌八氢番茄红素合成酶基因，如 *crt*B1、*crt*B2，以及编码后面将介绍的脱饱和步骤基因 *crt*I1、*crt*I2 和 *crt*I3[4]，尚未发现催化下一步向细菌红素转化的基因。在一条类胡萝卜素分支途径中，番茄红素被基因 *crt*Y 编码的酶环化形成 β 胡萝卜素，然后又被基因 *brp* 与 *blh* 编码的酶氧化裂解成视黄醛（图 4）[18]。在 NRC-1 菌株中的基因能参与类胡萝卜素生物合成途径中某些步骤，它们可以被光和氧气差异调控。

能量代谢

菌株 NRC-1 在好氧或厌氧条件下能进行光有机营养性生长，并有利用细菌视紫红质的光营养能力。它的生长需要 20 种氨基酸中的 15 种，有几种氨基酸可用作能源。有氧时，精氨酸和天冬氨酸能通过三羧酸循环被利用，厌氧时，精氨酸能够通过 pN-RC200 上 *arc*RACB 基因编码的脱亚胺酶途径被利用（图 4[3]）。靠氨基酸生长时，碳水化合物合成糖异生途径的基因和几乎所有反向 EM 糖酵解途径的基因都存在。尽管报道盐杆菌不能利用糖，在 NRC-1 中却有糖运转蛋白和葡萄糖脱氢酶、2-酮-3-脱氧葡萄糖酸激酶、半磷酸化 Entner-Doudoroff 途径的编码基因，还有负责糖异生和由 3 磷酸甘油醛（葡萄糖代谢产生）到丙酮酸代谢的基因。NRC-1 菌株还含有类似细菌的脂肪酸 β 氧化途径的酶和 2-酮酸脱氢酶复合体的编码基因。

进化和水平向基因转移

盐杆菌 NRC-1 是一种在进化上有意义的生物，它与一些产甲烷菌有较远的亲缘关系，根据 16SrRNA 序列分类为真古生菌（euryarchaeote）。用 DARWIN 软件对 NRC-1 与其他 11 种微生物的全基因组进行比较[4]，确证了 NRC-1 的古生菌地位，它与闪烁古生球菌（*Archeoglobus fulgidus*）和詹氏甲烷球菌（*M. jannaschii*）的亲缘关系最近。然而，它却与 G^+ 的枯草芽孢杆菌和放射性抗性菌耐放射异常球菌（*Deinococcus radiodurans*）也有一定相似性。最近，用多个已完成的基因组进行全基因组分析发现，NRC-1 由古生菌进化树根部分支（图 1）[19]。16S rRNA 与全基因组的进化树不相符需要进一步详细研究，这表明嗜盐生物可能出现在很早期的一个进化位点，然而，另一种可能是 NRC-1 在全基因组进化树中的位置被扭歪，由于许多细菌的基因水平转移，使盐杆菌属从其他古生菌中偏出而靠向细菌。

近来，对菌株 NRC-1 基因的来源又进行了深入研究（S.P.Kennedy 和 S.DasSarma，未发表）。在 NCBI 的 Clusters of Orthologous Groups 数据库中，对细菌进化有关的蛋白进行了种系进化分析，同时对簇居于基因组且编码特殊代谢途径的细菌类似基因也进行了种系进化分析，发现数百种蛋白，包括生物合成、转运、能量系统（如组氨酸利用、嘌呤代谢、甘油利用）和电子传递链组分都明显来源于细菌，这些基因好像是通过水平基因转移而被好盐生物获得的。令人惊奇的是，没有观察到这些细菌基因与 IS 元件的物理联系，表明在进化早期获得的这些基因连重组的痕迹已经消失。尽管域间生物遗传交换机制还不清楚，但是在 NRC-1 中发现的上百种细菌基因，反映出经过进化后在共栖于高盐环境的好盐细菌与古生菌之间有长期互换机会。NRC-1 基因组中拥有大量通过水平转移而获得的基因，这与其他嗜温古生菌[20]和超嗜热细菌[21]类似。

呼吸链组分的获得

有两类最有趣向菌株 NRC-1 进行水平基因转移的例子，即编码电子传递链因子和编码蛋白生物合成基因[11]。编码 NADH 脱氢酶大亚基的 10 个 *nuo* 基因，还有编码细胞色素 C 氧化酶亚基的 3 个 *cox* 基因，簇居在一个操纵子中，用于维生素 K2 类生物合成的 6 个 *men* 基因也是这样。有趣的是，*nuo* 基因的顺序同大肠杆菌一样，与其最相近的分支是蓝细菌集胞藻（*Synechocystis* sp.）PCC6803；*men* 基因的排序与大肠杆菌和耐放射异常球菌（*D*. *radioduran*）一样，最近的分支是枯草芽孢杆菌，而且这两组基因的 G+C 含量与染色体基因的平均值有差别（64 或 73%相对于 68%）。这些结果表明，嗜盐生物对氧化性大气的适应是通过基因水平转移，从好氧细菌中获得电子呼吸传递链而实现的。至于在现有嗜盐生物多样性进化中，呼吸基因的转移是一次发生还是多次发生，有必要进一步分析断定。

紫质膜的进化

含视黄醛的发色蛋白（如紫质膜中的细菌视紫红质）和感光视紫红质，近来已在多

种细菌和真核生物中发现，它们分布在生命所有三个分支中：古生菌、细菌和真核生物[12, 22]。尽管视黄醛发色蛋白的进化起源现在还不清楚，但是它们在自然界中的广泛分布与水平传播是一致的。进一步推测，原始视紫红质是早期进化结果，并且可能与海洋中最初占主导地位的光营养形式有关，早于靠叶绿素的光合作用。由于有相对简单跨膜光驱动质子泵来合成 ATP[22, 23]，这种海洋中早期光营养可能在还原性大气中就已经产生（尽管少量氧气对视黄醛的合成是必需的）。具有基于叶绿素、更复杂、运转高效光合作用系统的生物，随后在进化过程中替代大多数环境中的紫质膜生物。有趣的是，紫质膜吸光谱的峰值为 568nm，而光合作用膜在这个波段是波谷，它们的光谱互补性令人惊奇（图 5），这与两种膜的共进化关系一致。此外，基于叶绿素的蓝细菌和基于紫质膜的盐古生菌在现代高盐环境中仍然共存，只是前者在盐度低的环境中占优势，后者在饱和盐度中占优势。

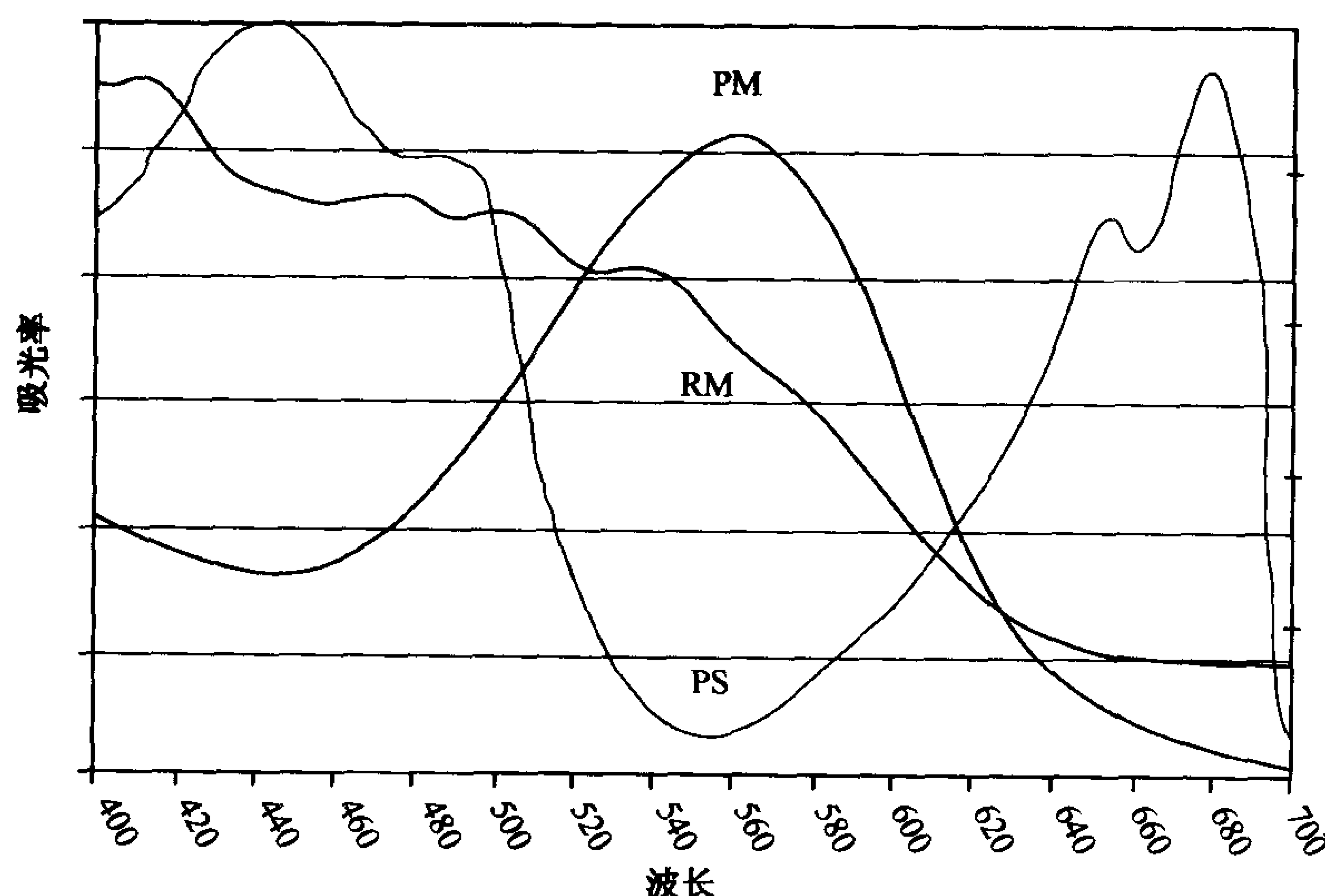


图 5　盐杆菌（*Halobacterium*）NRC-1 紫质膜（PM）、红膜（RM）和光合成膜（PS）的可见光谱。紫质膜和红膜用蔗糖梯度分离。紫质膜和光合成膜光谱的互补性很明显，这与两种膜的共进化相一致。

复制子 pNRC 的进化

NRC-1 基因组有一个大染色体和两个有亲缘关系的染色体外复制单元，这种基因组的组成复杂而又耐人寻味。保持多复制子（包括 pNRC100 与 pNRC200）的一个理由，是它们获取了一些必需的基因、因而对其生存是必要的。这些亲缘复制子间的相容性可以这样来解释：多个相容性不同的小组有多个复制起始子[3, 24]。在这些复制子上有数十个 IS 元件的拷贝，鉴于此，转座元件增加了复制子间 DNA 的频繁交换。此外，一旦染色体外复制子建立 2 个或 2 个以上拷贝，小复制子上的各个拷贝与大染色体之间的连续 DNA 交换，导致基因组更多多样性形成。

据推测 pNRC100 的进化机制是由 IS 元件介导了两个过程，包括前体质粒多复制子的融合和随后染色体基因的获得[3]。通过两个 IS 元件不对等的交换，pNRC100 前体复

制子某部分的复制导致反向重复序列的形成，随后有助于稳定重复序列内的区域并产生反向异构体。通过这种方法，必要基因会从染色体中获得并稳定在 pNRC100 和 pNRC200 复制子上，从而导致后二者微型染色体地位的确立。

菌株 NRC-1 基因组非常新的特征是含有多个微型染色体、能够获得新基因且能容纳多种必要基因，结果，NRC-1 基因组是多个必要复制子之间的一种竞争性动态平衡。这种状态在进化中可随时发生，并因复制子融合而数目减少，最终形成间歇期而沉寂。在盐杆菌属的菌株中，微型染色体的异质性就是这种基本动态过程的证据[25]，复制子间的竞争是进化中的一种普遍现象，并在打造原核基因组的长期进化中起重要作用，包括从质粒向新染色体的进化，鉴于已在菌株 NRC-1 中发现，这种推测是有道理的。

未来前景

NRC-1 全序列为极端嗜盐古生菌的进化和比较基因组分析提供了很好的平台[4, 11]，它作为几个已测序嗜中温古生菌之一，共栖于多种细菌繁衍的动态环境中，从基因组中发现了几百个细菌或来源未定的基因。开展更多各种嗜盐生物（如海洋盐古生菌）[1]基因组研究，对了解这些新型微生物的进化地位很有必要。不断变化的染色体外大复制子，包含着必需基因和大量 IS 元件，这些发现表明出现了参与基因竞争的多染色体[3]。

除了在进化上的洞察，嗜盐生物培养上简单和广泛的生物学感应，将为功能基因组和生物技术研究提供重要的发展机遇。DNA 点阵、蛋白质组和基因敲除都是进一步开展盐杆菌生物学研究良好的技术手段[2, 26]。近来，利用全基因组微阵列研究紫质膜的表达，已经证明了功能基因组手段的威力，并提醒我们在后基因组时代需要坚持已建立严格的遗传学实践[27, 28]。更重要的是，嗜盐古生菌为真核生物基础研究提供了一个最佳模型（如 DNA 复制、转录和翻译）。最后，嗜盐蛋白及其复合体中很多蛋白是极其新颖的，它们为生物技术提供了一个全新有远景的机遇，包括新疫苗和抗生素的开发[29, 30]。

致谢

在我的实验室开展盐古生菌基因组研究得到了国家科学基金的大力支持。感谢 Halobacterium Genome Consortium 中的各位现在和以前的学生、同事和合作者，是他们提供了本章中收集的大量信息。特别感谢 Philip Harriman 博士的支持和鼓励。

（邵宗泽 译）

参考文献

1. DasSarma S, Arora P. Halophiles. In: Encyclopedia of Life Sciences. London: Macmillan, 2000, pp. 458–466.
2. DasSarma S, Robb FT, Place AR, et al. (eds). Archaea: A Laboratory Manual—Halophiles. Cold Spring Harbor, NY: Cold Spring Harbor, Laboratory Press, 1995.
3. Ng W-L, Ciufo SA, Smith TM, et al. Snapshot of a large dynamic replicon from a halophilic archaeon: megaplasmid or minichromosome? Genome Res 1998; 8:1131–1141.
4. Ng WV, Kennedy SP, Mahairas GG, et al. Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci USA 2000; 97:12,176–12,181.
5. Joshi JG, Guild WR, Handler P. The presence of two species of DNA in some halobacteria. J Mol Biol 1963; 6:34–38.
6. Charlebois RL, Doolittle WF. Transposable elements and genome structure in halobacteria. In: Berg DE, Howe MM (eds). Mobile DNA. Washington, DC: American Society for Microbiology, 1989, pp. 297–307.
7. Ng W-L, Kothakota S, DasSarma S. Structure of the large gas vesicle plasmid in *Halobacterium halobium*: inversion isomers, inverted repeats, and insertion sequences. J Bacteriol 1991; 173: 1958–1964.
8. Hackett NR, Bobovnikova Y, Heyrovska N. Conservation of chromosomal arrangement among three strains of the genetically unstable archaeon *Halobacterium* species. J Bacteriol 1994; 176: 7711–7718.
9. St Jean A, Charlebois RL. Comparative genomic analysis of the *Haloferax volcanii* DS2 and *Halobacterium* sp GRB contig maps reveals extensive rearrangement. J Bacteriol 1996; 178: 3860–3868.
10. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res 1998; 8:195–202.
11. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S. Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res 2001; 11:1641–1650.
12. DasSarma S, Kennedy SP, Berquist B, et al. Genomic perspective on the photobiology of *Halobacterium* species NRC-1, a phototrophic, phototactic, and UV-tolerant haloarchaeon. Photosyn Res 2001; 70:3–17.
13. Baliga NS, Goo YA, Ng WV, Hood L, Daniels CJ, DasSarma S. Is gene expression in *Halobacterium* NRC-1 regulated by multiple TBP and TFB transcription factors? Mol Microbiol 2000; 36:1184–1185.
14. Baliga NS, Kennedy SP, Ng WV, Hood L, DasSarma S. Genomic and genetic dissection of an archaeal regulon. Proc Natl Acad Sci USA 2001; 98:2521–2525.
15. Bolhuis A. Protein transport in the halophilic archaeon *Halobacterium* sp NRC-1: a major role for the twin-arginine translocation pathway? Microbiology 2002; 148:3335–3346.
16. Patenge N, Berendes A, Engelhardt H, Schuster SC, Oesterhelt D. The *fla* gene cluster is involved in the biogenesis of flagella in *Halobacterium*. Mol Microbiol 2001; 41:653-663.
17. DasSarma S, Arora P. Genetic analysis of the gas vesicle gene cluster in haloarchaea. FEMS Microbiol Lett 1997; 153:1–10.
18. Peck RF, Echavarri-Erasun C, Johnson EA, et al. *brp* and *blh* are required for synthesis of the retinal cofactor of bacteriorhodopsin in *Halobacterium*. *J Biol Chem* 2001; 276:5739–5744.
19. Korbel JO, Snel B, Huynen MA, Bork P. SHOT: a web server for the construction of genome

phylogenies. Trends Genet 2002; 18:158–162.

20. Deppenmeier U, Johann A, Hartsch T, et al. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol 2002; 4:453–461.
21. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 1999; 399:323–329.
22. Beja O, Aravind L, Koonin EV, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 2000; 289:1902–1906.
23. Racker E, Stoeckenius W. Reconstitution of purple membrane vesicles catalyzing light-driven proton uptake and adenosine triphosphate formation. J Biol Chem 1974; 249:662–663.
24. Ng WL, DasSarma S. Minimal replication origin of the 200-kilobase *Halobacterium* plasmid pNRC100. J Bacteriol 1993; 175:4584–4596.
25. Ng W-L, Arora P, DasSarma S. Large deletions in class III gas-vesicles deficient mutants of *Halobacterium*. Sys Appl Microbiol 1994; 16:560-568.
26. Peck RF, DasSarma S, Krebs MP. Homologous gene knockout in the archaeon *Halobacterium* with *ura*3 as a counterselectable marker. Mol Microbiol 2000; 35:667–676.
27. Baliga NS, Pan M, Goo YA, et al. Coordinate regulation of energy transduction modules in *Halobacterium* sp analyzed by a global systems approach. Proc Natl Acad Sci USA 2003; 99: 14,913–14,918.
28. DasSarma S. Biology reports Ltd. faculty of 1000 commentary. Available at: http://www.faculty of1000.com/article/12403819. Accessed January 8, 2003.
29. Stuart ES, Morshed F, Sremac M, DasSarma S. Antigen presentation using novel particulate organelles from halophilic archaea. J Biotechnol 2001; 88:119–128.
30. Hansen JL, Ippolito JA, Ban N, Nissen P, Moore PB, Steitz TA. The structures of four macrolide antibiotics bound to the large ribosomal subunit. Mol Cell 2002; 10:117–128.

第六部分：基因组数据库的应用

22

微阵列表达分析和细菌基因组

Carsten Rosenow and Brian Tjaden

引言

基因组信息的快速积累和同步发展的微阵列技术，为研究者们提供了一个机会，使他们能以前所未有的速度研究生物复杂性。大量已测序基因组可用来发展多种用途的微阵列，这些用途包括 DNA 表达分析、比较基因组杂交、大规模基因分型（genotyping)、转录组（transcriptome）分析，以及 RNA 衰减的研究。此外，微阵列技术具有加速药物开发进程的潜能，并能在诊断市场中扮演重要角色（图 1)，新技术的发展与高通量技术和强大计算功能相结合是基因组时代的主要进步之一。

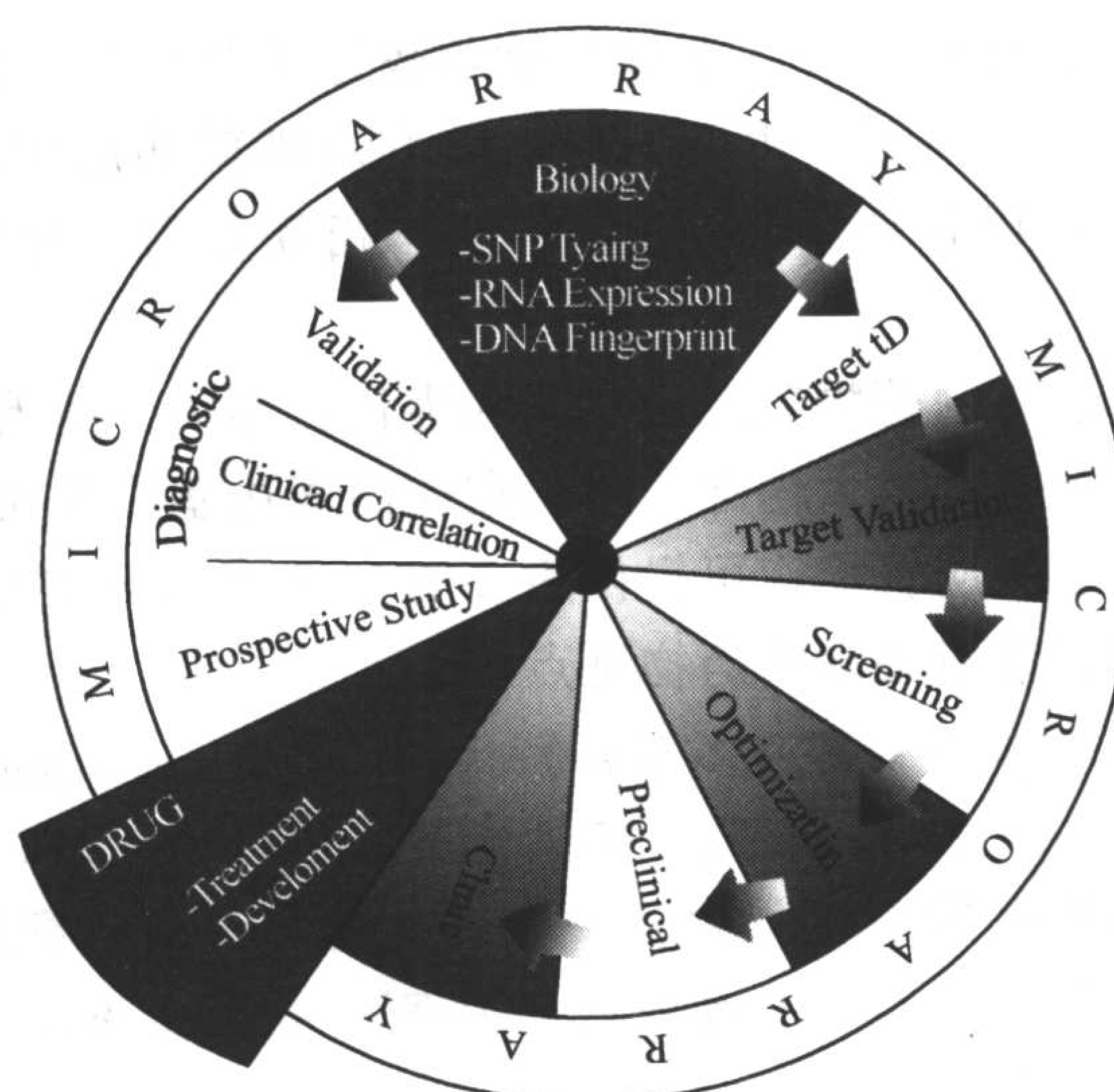


图 1　微阵列的应用对药物开发过程中所有领域的影响，对诊断市场具有潜力。

技术

微阵列是由脱氧核糖核酸（DNA）与玻璃等固相支持物连接组成，DNA 可以是 PCR 产物、cDNA（互补 DNA）转录物、或与待分析靶序列互补的寡核苷酸。由于这些微阵列可实现高通量，因而能获得大量的平行数据，此外，生物信息学和数据分析先进的应用软件，可将这些海量数据翻译成有意义的结果（在数据分析一节将详细描述)。

目前，有两种微阵列技术：点制微阵列和寡核苷酸探针阵列。点制微阵列采用预先

获得的 cDNA 或聚合酶链反应产物，将它们点在玻璃表面。寡核苷酸探针阵列是在玻璃表面原位合成序列已知的一系列寡核苷酸，寡核苷酸合成是通过喷墨打印或与半导体工业类似的光刻印刷方法来完成。然而，由于点制技术和原位合成技术间存在根本差异，因此在可靠性、密度和重现性方面表现出不同的水平，二者获得的表达分析结果通常不具可比性。要充分了解这两种技术，请参阅近期的综述[1~3]。

表达分析

基因组学正在改变对生物的理解，DNA 测序项目给生物有机组成提供了一个整体认识，然而，要想了解细胞的进程，除了基因和基因组知识外，还必须了解基因的功能和基因间的相互作用。微阵列技术可用于同时检测生物体内大量 mRNA 的表达水平，在这项应用中，cRNA[4]、RNA 或 cDNA[5,6]被荧光标记，并与微阵列进行杂交。根据标记技术的不同，RNA 表达水平要么以参照样品为基准用相对值表示（通常在点制阵列中），要么以绝对值表示 RNA 在给定细胞中的表达水平（通常在高密度寡核苷酸探针阵列中）。通过发展 RNA 直接标记技术和 cDNA 合成方法，克服了原核生物 RNA 表达研究的初步障碍[5,7]。

早期研究着眼于在不同生长条件下微生物的表达差异[8,9]、细菌对药物的反应[10]和环境改变对基因表达的影响[6]。用不同的微阵列对细菌病原体和受感染人细胞的表达差异进行同步研究，可以帮助了解寄主-病原体相互作用，并导致新药物靶点的发现[11,12]。

在微生物中最显而易见的药物靶点是对体内生长必不可少的蛋白质，或是与其毒性相关的蛋白质，用微阵列技术大大增强了对只在体内表达基因或在感染过程中表达基因的确认。此外，了解寄主的应答反应可提供更多寄主细胞表面的药物靶点（黏附分子）或细胞内对感染或细菌存活所必需的药物靶点。

Cohen 等[13]鉴定出在单核细胞增多李斯特氏菌（*Listeria monocytogenes*）感染人早幼粒细胞系（promyelocytic cell line）后，表达明显不同的 97 个人类基因，并强调，这些研究不仅加强了对病原体致病机制和分子生理的理解，而且帮助确认了新的治疗靶点。鉴于近来由于多种药物抗性机制和抗药性病原体的数量不断增加而引起的危机形势，这一点显得尤为重要[14,15]。

微阵列行业最重要的短期任务之一是建立公共数据库，以微阵列实验为基础构建的代谢和调控途径、定义转录调控因子、模拟细胞过程和相互作用。为实现这个目标，收集并提供原始微阵列数据十分重要。一些大型研究所已建立了公共数据库[16~19]（www.ncbi.nlm.nih.gov/geo），此外，其他数据发表在同行杂志的网页上或在作者的个人网页上。

MIAME 计划 1（微阵列试验的最少信息，Minimum Information About Microarray Experiment）推动了微阵列数据的标准化，以及试验细节的公布[17]，这项计划规定保证微阵列数据很容易解释，微阵列结果可独立验证所需要的最少信息。

功能基因组领域的进步，部分决定于科学家们公开他们的数据上，如果与其他分析手段相结合，表达数据库将具有更大的价值，这些其他分析手段包括：聚类分析、比较

基因组杂交、DNA 序列初步注释、蛋白质组资料或与其他生物体的同源性鉴定（图 2）。

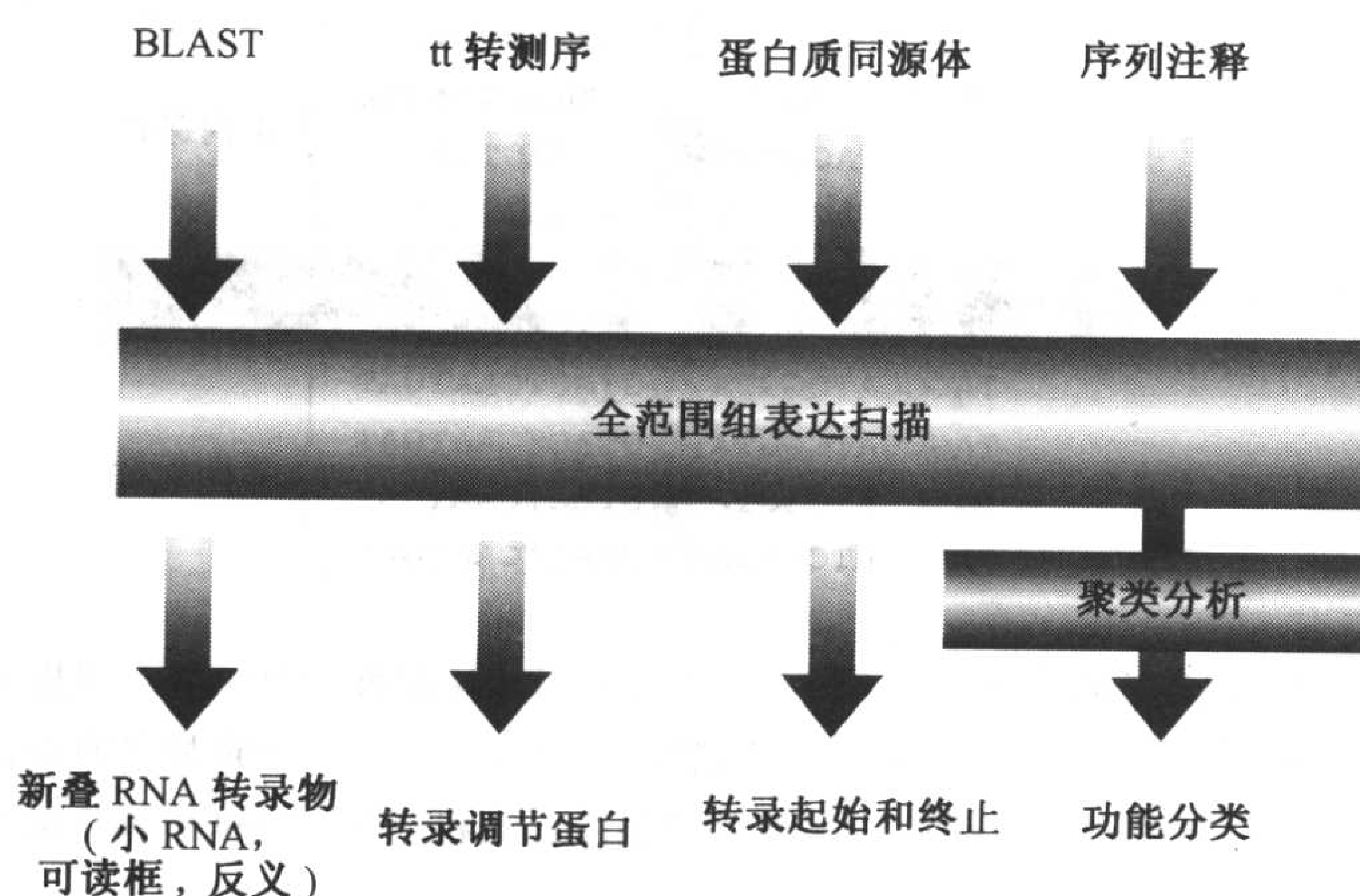


图 2　全基因组表达扫描与数据库信息和序列分析相结合，可加速基因特性研究和发现细胞内新途径和新调控机制。

用微阵列进行比较基因组杂交、再测序和基因分型

比较基因组杂交和遗传多样性研究，对了解微生物致病机制和追踪传染病爆发的重要性不断提高，高密度寡核苷酸探针阵列使对任何生物的快速比较基因组杂交、再测序和随后的基因分型成为可能。再测序在确定不同表型特征菌株间的基因型差异研究中已越来越重要，为了用微阵列确定特定位点上的序列与已知序列的差异，在微阵列上滴定 4 种寡核苷酸探针，这 4 种探针仅在中间位置碱基或第 13 位碱基上不同，每种探针包含 4 种可能核苷酸中的一种（图 3）。为了增加冗余度，每个靶序列的正链和负链都可检测，根据所用微阵列的密度和尺寸，长至 120kb 的连续或不连续序列信息可通过一次杂交获得，这些微阵列数据的平均灵敏度 80％以上，准确率达 99.999％[20]。

对无需特异核苷酸序列的应用，可通过比较基因组杂交获得对基因组更全面的认识，基因组的比较可以帮助更好地理解细菌菌株的进化、病原体和非病原体之间的差异和确定细菌菌株间或亚型间基因的不同。所有这些信息与全基因组表达分析一起可提供基因型和表现型的信息，这些信息更好地了解两种相近菌株致病性之间的根本不同，促进基因特性和功能的研究（功能基因组学）。

这种方法的例证是 Kato-Maede 等，利用结核分枝杆菌（*Mycobacterium tuberculosis*）高密度寡核苷酸阵列进行的研究[21]，他们在 19 个临床和传染性特征明显的结核分枝杆菌分离株中检测一系列基因缺失，这些缺失存在于所有分离株中，但在不同分离株中缺失的程度不同，这些资料显示，与人类线粒体 DNA 中的变化类似，基因组缺失可用于重建进化树。

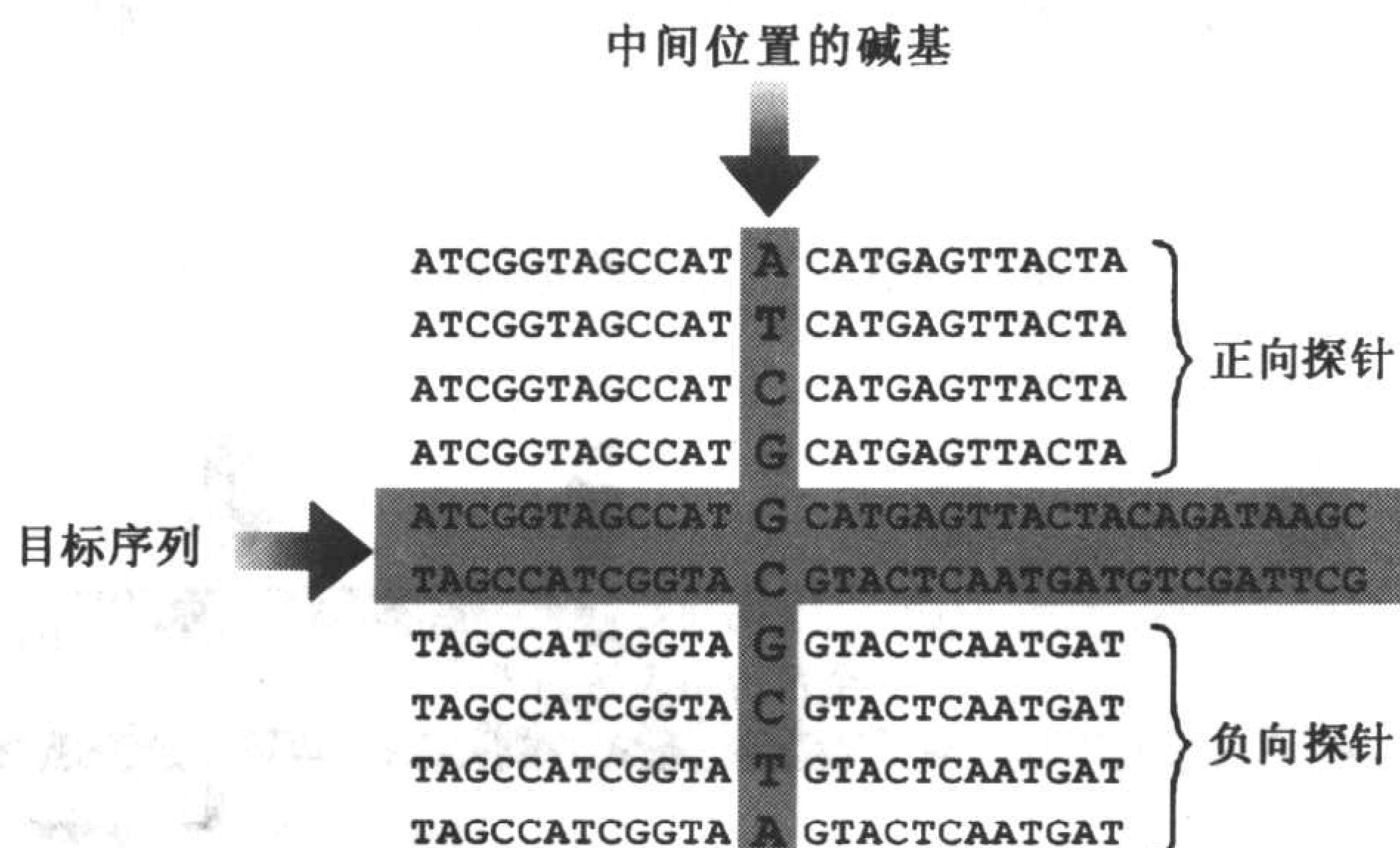


图 3　比较序列分析阵列的设计原则。序列分析采用 4 探针探查策略，即用 4 条 25 聚体寡核苷酸探针确定序列中间位置碱基类型。在 4 条探针中最稳定的杂交体将产生最强的荧光信号。一种客观的统计算法对正链和负链中的每种碱基赋予特性值。

比较基因组杂交的其他应用，包括鉴定细菌不同分离株中十分普遍的染色体重排[9,22~24]，这些重排由同源重组产生，包括缺失、复制和倒位。此外，脑膜炎奈瑟氏菌（*Neisseria meningitidis*）基因组中的重复因子，导致了抗原的多样性，并通过促进重组导致基因组的流动性（fluidity）[25,26]。

细菌种类遗传多样性的另一原因是基因的水平转移，这是适应新环境很重要的步骤[27]。基因水平转移也可用微阵列技术进行研究。在金黄色葡萄球菌（*Staphylococcus aureus*）中，包括抗生素抗性基因在内的许多毒力相关因子由水平转移获得[28,29]。用 DNA 微阵列比较 5 种不同非致病大肠杆菌与大肠杆菌菌株 MG1655 的基因组序列，证实在整个染色体中存在高度多样性[7]，令人惊奇的是，鉴定出的变异区由非常多假定可读框和移动遗传因子组成[30]。另一个研究小组利用 DNA 阵列进行了肺炎链球菌（*Streptococcus pneumoniae*）比较基因组杂交，获得了相似结果，他们总结认为可变基因组区域，通常与转座子或噬菌体有关[31]。

结合计算机模拟（*in silico*）分析和比较基因组杂交，可确定生物体的遗传异质性，并可为微生物学实验室提供一个强大的工具[32]。

数据分析

微阵列实验室的最大优点是能够同时分析大量基因的表达水平，这就为科学家们对成组基因的研究提供了有力的工具。然而，科学家一定要慎重处理从微阵列实验中获得单个基因的表达数据，生物学是复杂的，微阵列实验有不确定性、误差和噪音。细胞生长、mRNA 提取、荧光标记的不精确步骤，都可能导致结果偏差，杂交是一个随机过程，在微阵列实验中交叉杂交和非特异杂交结果必须控制到最小，况且，微阵列的激光扫描在确定每个点的强度大小及其背景的同时，也会将噪音引入到表达数据中。

尽管技术的不断提高减少了这些误差的来源，大多数研究者仍竭力提倡重复试验以提高结果的可靠性。当然，最好是收集在不同生长条件下的细胞表达数据，探查表达谱

(expression profile) 的最佳范围，因此，对经费固定的微阵列实验，研究者必须决定如何平衡预算，一方面尽量在相同条件下进行重复试验，以提供更多有统计学意义的结果，另一方面要争取在不同试验条件中寻找最佳表达范围。

根据引起误差的原因重复测定，可在微阵列实验过程中的任何阶段进行，通常情况下，阵列中的每个基因含有多个点，并允许在同一微阵列芯片上进行重复测定和重复试验。举一个例子，图 4 显示了流感嗜血菌（*Haemophilus influenzae*）转录表达分析 cDNA 阵列中的一部分，该阵列中每个基因有 4 个完全一样的点。



图 4　流感嗜血菌（*Haemophilus influenzae*）cDNA 阵列中的一部分。每一个基因有 4 个完全一样的点。

在其他微阵列，如寡核苷酸阵列，每个基因可包含多条探针，这些探针能分析基因内不同区段的转录。图 5 显示了大肠杆菌基因 *rpsB* 的表达水平，该阵列中包含 15 种不同寡核苷酸探针，为了从单次实验数据中估计基因表达情况，简单的解决方法是计算 15 种探针表达水平的均值，采用这种策略，每种探针对平均表达水平的贡献相同。

然而，用寡核苷酸探针判断基因表达的挑战，是不同寡核苷酸具有不同杂交亲和力，因此，用不同探针分析相同转录时，常常得到明显不同的表达水平。如果有多组实验数据，则根据探针的亲和力对它们进行不同程度的加权，这样，基因表达水平就是 15 种探针的加权平均值（图 5）。

通过最大期望值法（expectation maximization，EM），捕获每条探针的独特响应值，对基因的表达水平提供更精确的估计[33]，最大期望值法反复估算 4 次实验中每次基因 *rpsB* 的表达水平（期望步骤），在这些估算表达水平的基础上，计算代表每条探针独特

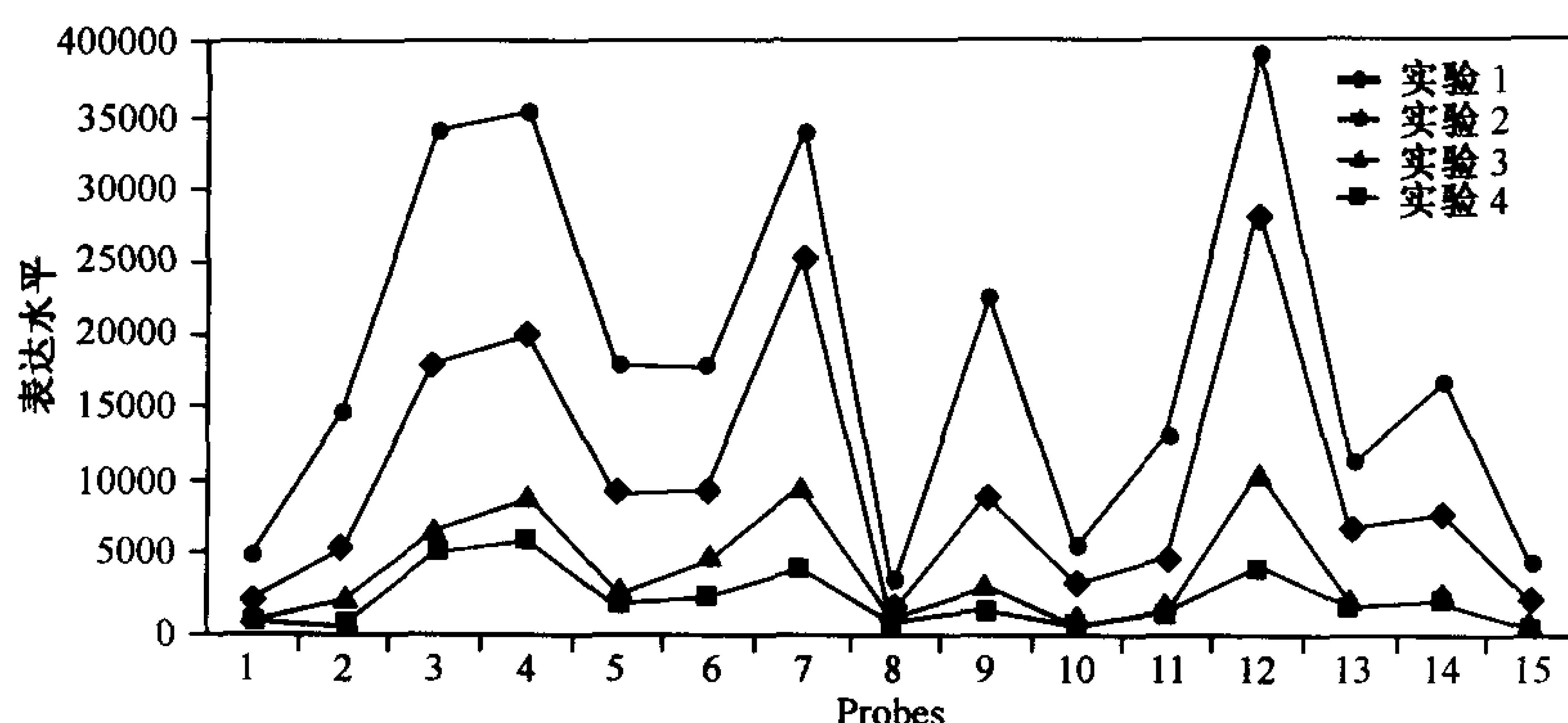


图 5　用寡核苷酸阵列从多次试验中计算大肠杆菌基因 *rpsB* 表达水平。每次实验都包含基因 *rpsB* 的 15 条探针。

响应值的一组参数的优化值（最大化步骤）。如果不是用多个寡核苷酸探针，而是用多次实验检测同一基因的表达，则需要用多个重复阵列，或是在一个阵列中含有寡核苷酸重复点或 cDNA 重复点，这时最大期望值法也要做相应调整。在最大化步骤中就不再计算探针响应度的优化值了，而是计算背景和单个点（或寡核苷酸）误差的优化值。通过每个基因的多点重复或每种条件下的微阵列重复，基因表达水平的统计显著性可以定量，并最终获得更加可靠的结果[34,35]。

每次试验中每个基因表达水平一旦计算出来，表达水平数据通常以矩阵形式列出，如图 6 所示，矩阵中每列对应一个基因，每行对应一次试验，如果在 m 次实验中分析 n 个基因，表达矩阵即为 $n \times m$，这时，可以根据表达谱研究基因的相似性。目前有多种相似矩阵或距离矩阵，如皮尔逊相关（Pearson correlation）、欧几里得距离（Euclidean distance）、卡方检验（chi square）、斯皮尔曼等级相关（Spearman rank correlation）和 Kendall' s tau。图 7 举例了其中最常用的皮尔逊相关系数，矩阵中包含两个基因 a 和 b，a 是列向量（a_1，a_2，$...a_m$），b 是列向量（b_1，b_2，$...b_m$），因此，两个基因（或两次试验）表达向量的相关系数 r 可用这种方法计算出来，如果要计算所有基因对的相关系数，可以用 $n \times n$ 相关系数矩阵，矩阵中每个值都代表两个基因表达的相似性。用这种方法，对任何给定的基因，在实验中都会发现有相似表达的其他基因。

	基因 1	基因 2	基因 3…		基因
实验 1	150	211	478	…	982
实验 2	817	300	525	…	700
…	…	…	…	…	…
实验 *m*	073	009	112	…	213

图 6　*n* 个基因的 *m* 次实验表达水平数据矩阵示例。矩阵中的每列对应特定基因在 *m* 次实验中每次的表达水平，每行对应特定试验中 *n* 个基因的表达水平。

$$r=\frac{m\sum_{j}a_{i}b_{i}-\sum_{j}a_{i}\sum_{j}b_{i}}{\sqrt{\left[m\sum_{j}a_{i}^{2}-\left(\sum_{j}a_{i}\right)^{2}\right]\times\left[m\sum_{j}b_{i}^{2}-\left(\sum_{j}b_{i}\right)^{2}\right]}}$$

图 7 对两个表达向量分别为（a_1，a_2，$\cdots a_m$）和（b_1，b_2，$\cdots b_m$）的基因，上图显示 m 次实验中皮尔逊相关系数的计算公式。相关系数 r 的范围在 $-1 \leqslant 0 \leqslant 1$ 之间。当 r 值接近 1 时表示基因表达高度相关，当 r 值接近 0 时表示基因表达无相关性，当 r 值接近 -1 时表示基因表达为负相关。

当相关系数矩阵计算出来后，可通过聚类方法进一步分析表达相似的基因群，数据点（基因的表达向量）位于通常很大的 m 维空间内，实验次数不同则 m 值不同，从计算角度看，表达谱聚类不那么简单。有许多不同运算法用于基因表达的聚类分析，包括层级聚类（hierarchical clustering）[36]、逐步聚类分析（k mean）[37]、聚类亲和搜寻技术(Cluster Affinity Search Technique，CAST)[38]、自组图（self-organizing maps）[39]和基于模型聚类（model-based clustering）[40]，随后介绍几个常用策略。

层级聚类是一种“贪心的”聚类策略，其优点是简单，并将最终聚类结果表现为容易浏览的层级树图，如何确定两群数据点之间的距离，可以改变层级聚类的分析结果。在单连接聚类（single-link cluster）中，两群之间的距离是一群中任何点与另一群中任何点之间的最短距离。在平均连接聚类（average-link cluster）中，两群之间的距离是两群质心之间的距离。在完全连接聚类（complete-link cluster）中，两群之间的距离是一群中任何点与另一群中任何点之间的最远距离。不同的距离矩阵需要不同的计算量，并影响最终的聚类层级图式。

1. 让每个表达向量成为包含一个数据点的群。

2. 找到具有最小距离的两群（单连接、平均连接和完全连接），并将它们合并为一个群。

3. 重复步骤 2 直到所有群合并为一个群（即层级树的根）。

k mean 是一种常用并且相对简单的启发式方法，将数据点聚类到 k 群，这实际上是最大期望值法应用到混合密度的特例，其中混合成分呈高斯分布[41]。

1. 随机赋予每个点一个介于 1 和 k 之间的聚类值。

2. 对每一个 k 群，计算赋予相同聚类值所有点的平均值。

3. 对每一个数据点，确定该点与每个 k 群平均值的距离，赋予该点一个最接近 k 群的聚类值。

4. 重复步骤 2 和 3 直到没有数据点可以被赋予与上一次重复不同的聚类值。

CAST 模拟数据点作为图表的顶点，在图表中每个顶点对之间有一个边表示它们表达向量的相似性，然后 CAST 搜索图表中的顶点并将它们分群，每个群中的每个顶点与该群中其他顶点充分相似，与 k mean 法相同，CAST 有一个阈值，该值限定了将数据点集聚成多少群。

1. 选择一个非聚类点，将其本身设为一群。

2. 如果任何非聚类点到群内所有点之间距离的平均值大于某个阈值，将非聚类点加入到该群中。

3. 如果任何群内任何点到群内其他点之间距离的平均值低于某个阈值，将该点移出到该群外。

4. 重复步骤 2 和 3 直到聚类稳定。

5. 将群内所有点标记为已聚类，如果仍有非聚类点存在，回到步骤 1。

表达数据一旦聚类，就有许多方法对数据进行分析，并获得有生物学意义的结果。聚类数据很直观，研究者可以检验有相似调控规律的基因或存在于同一细胞中的基因是否类聚在一起。如果未知功能基因与已知功能基因分在一群，则可以设法推断一些未知基因的功能。

然而，聚类分析也有它的局限性，尽管聚类分析可以获取表达谱的一些明显趋势，但是，有许多基因的表达谱不适合聚类分析中的任何群，因此导致集群上的误导。还有，大多数聚类分析技术要求使用者规定一些阈值参数来确定群的数目，通常情况下，由于大多数基因参与了多个细胞途径，没有一个好办法可以决定到底需要设定多少群。

将聚类分析与序列分析相结合，是对表达数据进行聚类分析最有希望的新领域，例如，假定表达相似的基因具有相似的调控机制，如果几个基因的上游 DNA 区段的表达谱聚类在一起，则可以通过序列搜索或局部排序来寻找这些基因共同转录因子的结合位点[37]。在这些表达分析方法进一步完善之前，这些分析方法主要用于对大量基因群进行功能性预测，并提供一个细胞途径的整体认识，而不是精确分析单个基因的表达（图 2)。

全基因组转录分析

正如本章所讨论的，微阵列可用于基因组序列已知有机体的全基因表达水平分析，然而，除了确定已知基因的表达外，微阵列已越来越多地用于定位新基因[42,43]、研究 RNA 衰减[44]和识别非翻译转录因子[42,43,45]。有机体的基因组序列一旦建立，即可设计寡核苷酸探针来分析整个基因组的表达，而不仅仅是那些最终被翻译的序列。

例如，在新微生物基因组序列测定后，第一步是利用计算机进行序列分析，并对基因进行注释。基因预测程序通常很准确，特别是识别基因编码序列，但它们却不擅长识别那些位于起始密码子上游，或终止密码子下游被转录但不被翻译的序列，分析全基因组表达水平的寡核苷酸微阵列，有助于识别这些非翻译区，从而确定转录的起始点和终止点。

图 8 显示了在 24 次微阵列实验中测定基因 *pfkA* 的表达水平，基因 *pfkA* 在大肠杆菌中编码磷酸果糖激酶，图中显示了分别对应编码序列和终止密码子下游探针的信号强度。下游序列的表达显示了 3′-非编码区和转录终止位置，对基因非翻译转录区的了解有助于认识和理解基因调控因子的作用。

多个基因转录成一个 mRNA 称为多顺反子、多基因或操纵子。当基因表达表明转录从一个基因编码序列延伸至相邻基因的编码序列时，这两个基因很可能产生一个多顺反子 mRNA。对于操纵子基因，基因间的区段也被转录，这可以用微阵列实验鉴定，

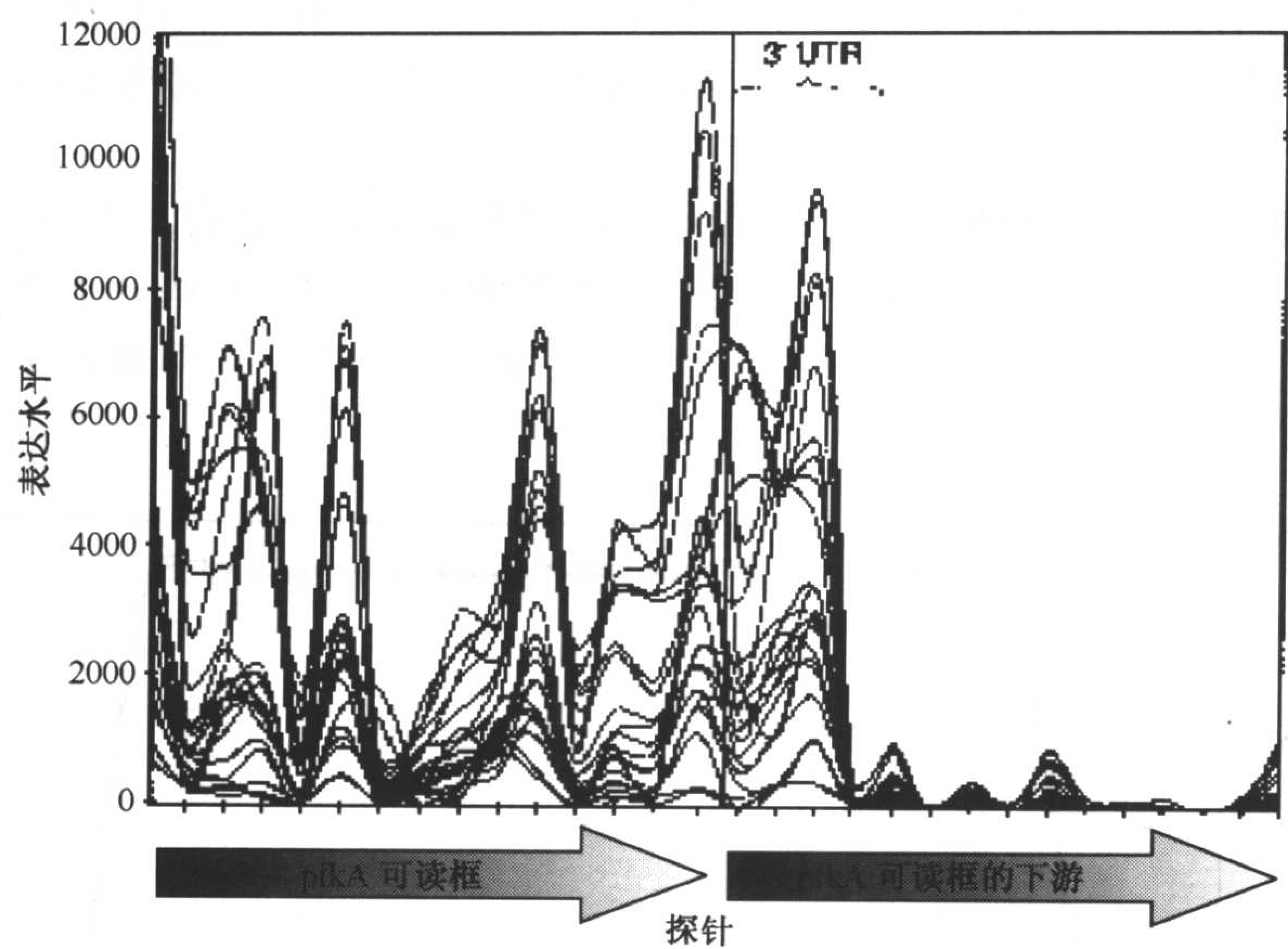


图 8 在 24 次微阵列实验中，测定的大肠杆菌磷酸果糖激酶基因 *pfkA* 编码序列和下游非翻译区的表达水平。

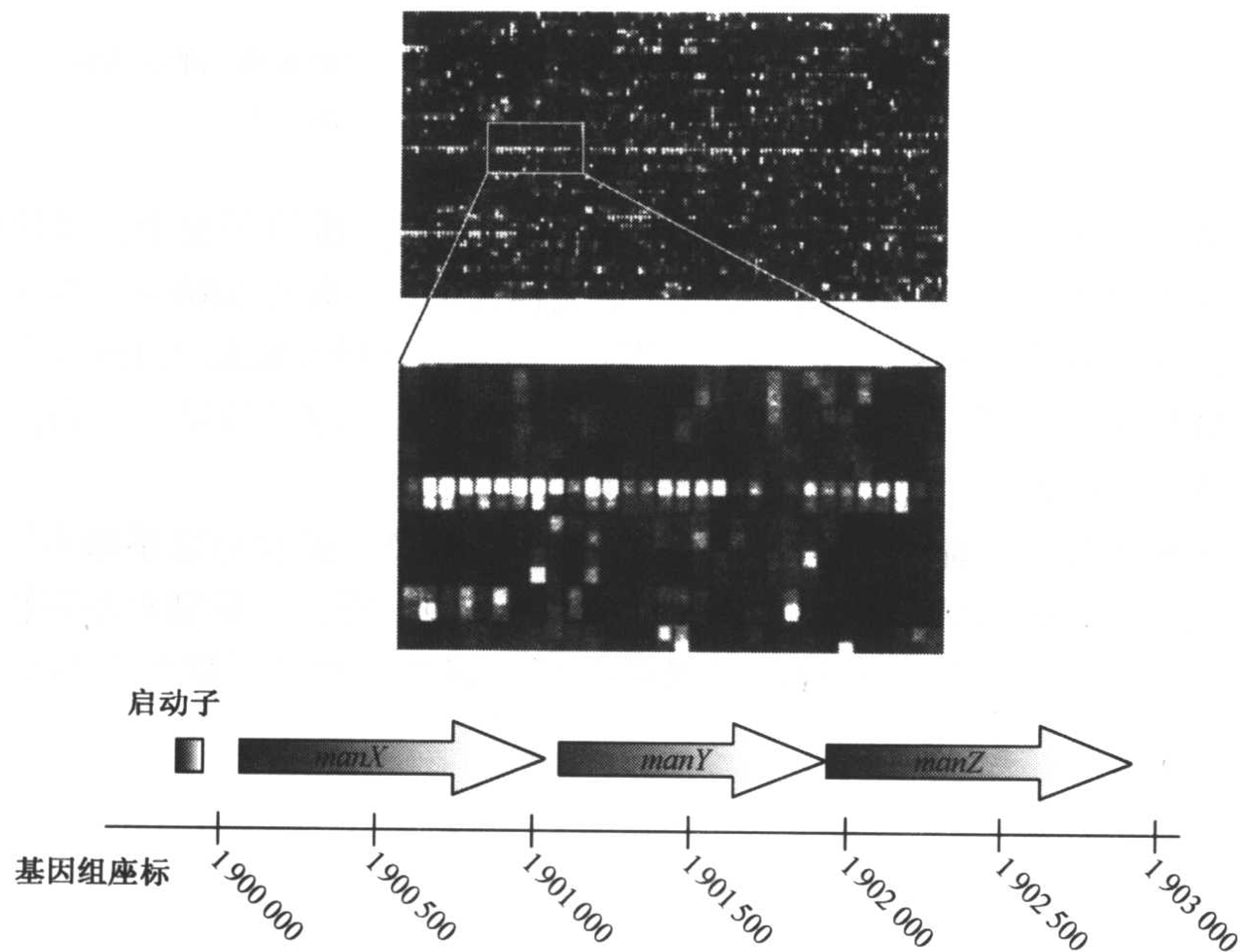


图 9 用寡核苷酸微阵列分析基因 *manX*、*manY* 和 *manZ* 的表达。这 3 个基因组成了参与甘露糖磷酸转移酶系统的操纵子。阵列中检测以上基因的寡核苷酸探针，按照大肠杆菌基因组中的相邻顺序排布。

图 9 显示了分析大肠杆菌 *manXYZ* 操纵子探针的微阵列杂交图像的信号强度[46]，该操纵子由 3 个基因（*manX*、*manY*、*manZ*）组成，为了使相邻基因的相似表达显而易

见，微阵列中包含了检测相邻基因的邻近探针。显然，共转录基因在转录水平上是共同调控的，同一操纵子中多个基因的功能也紧密相关，因此，鉴定哪些基因是多顺反子mRNA中的一部分，对理解基因的功能特别有用。

除了已知基因周围的转录区，微阵列还可以用来鉴定新转录物，如小RNA分子是短转录物，不被翻译，通常具有调控其他基因表达的功能[47]。这些分子常常在特定条件下表达，如微生物细胞稳定生长后期，图10显示了在3次不同微阵列实验中大肠杆菌基因 *csrB* 的RNA表达情况[43]。

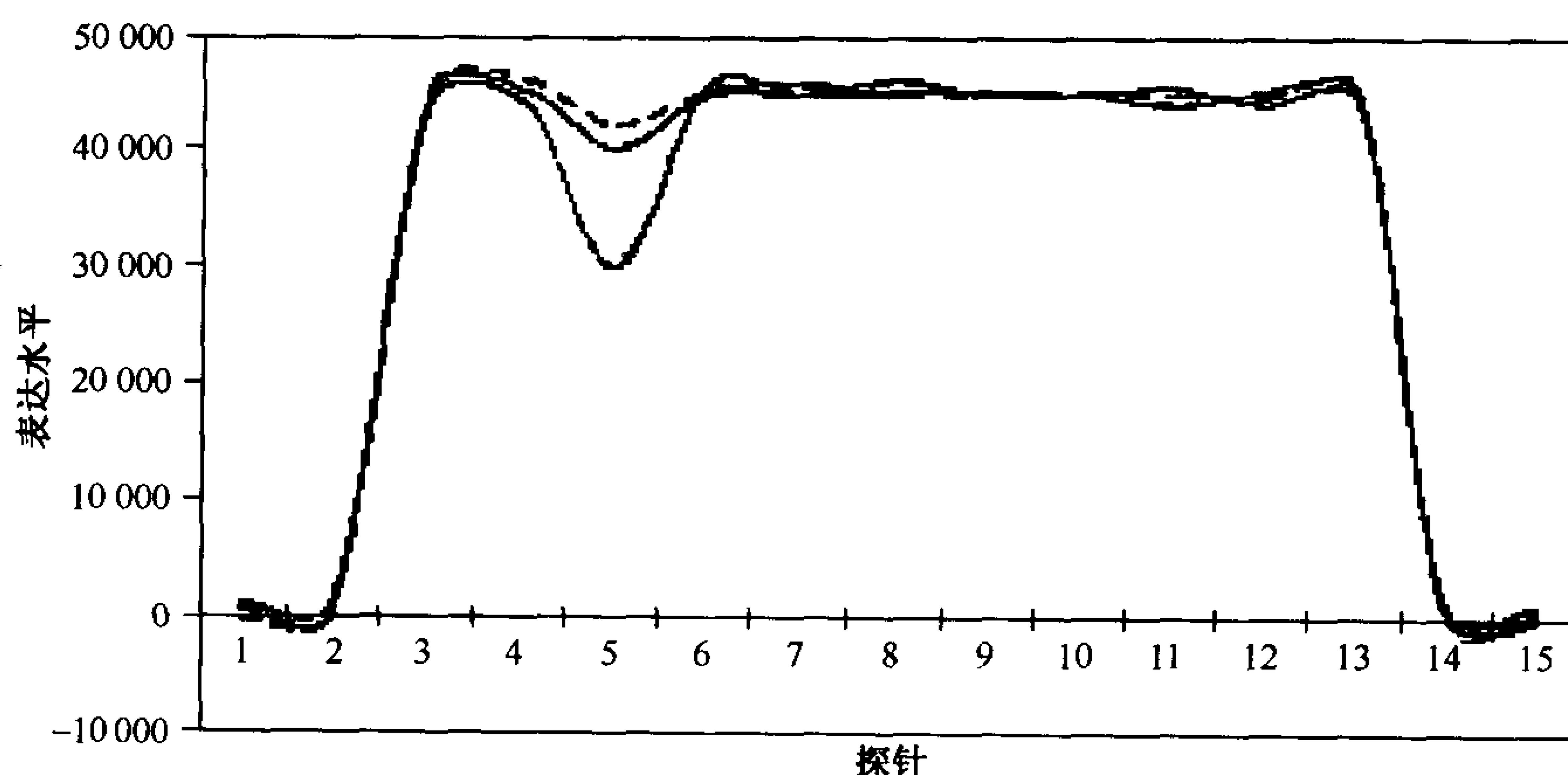


图10　寡核苷酸微阵列3次试验中大肠杆菌基因b2793和b2792间隔区的表达情况。探针3到探针13的高表达水平对应360bp的非翻译小RNA *csrB*。

仅有少数几个小RNA进行过实验研究，在初级序列分析的基础上，对这些分子进行识别非常具有挑战性，因为它们的功能在很大程度上由高级结构决定。如果从微阵列实验中检测到由未知功能序列产生的转录产物，那么该序列可能是小RNA分子编码区或是一个未知的可读框[42]，对基因组中那些新转录物区域的初级序列分析，可以预测转录物的功能和这些转录物是否值得进一步验证。

这只是微阵列强大功能的一个例子，一次简单的微阵列实验可提供微生物在给定生长条件下的全部转录表达情况，通过对已知基因表达的分析，为识别新基因提供预测指导，微阵列推动了对急速增长基因组资料的分析，有助于更多了解微生物的细胞机制(图3)。

结论

微阵列对基础研究、靶点鉴定、药物开发和诊断领域中的传统技术提出了挑战，将同时获得的大量数据与计算方法的进步相结合，可以把这些数据翻译成有生物意义的信息，这将促进此项技术在药物开发和诊断市场中的应用（图1）。对微生物感染期寄主和病原体的表达进行平行分析，可能发现人体和细菌细胞内新的重要新陈代谢途径和从未被发现的药物靶点。

除非下游技术的能力有了相应提升，这些靶点将储存在药物开发公司的数据库中，

在随后的药物筛选和临床验证阶段，用微阵列技术将使这些公司尽快跟上大量药物靶点的发现。

创造价值链的下一步是体外毒理学，把高通量技术与对药物的吸收、分布、代谢、排泄/毒性的计算机模拟相结合，更确切地了解药物特性，从而加速临床前的研究。把已鉴定的生理标记以及与已知药物的反应特性整理成数据库，最终导致更有效的预测药物效果和获得更可靠的药物剂量信息，同时又避免了不良的毒副作用。

（张 琼 译）

参考文献

1. Harrington CA, Rosenow C, Retief J. Monitoring gene expression using DNA microarrays. Curr Opin Microbiol 2000; 3:285–291.
2. Lipshutz RJ, et al. High density synthetic oligonucleotide arrays. Nat Genet 1999; 21(1 Suppl): 20–24.
3. Gershon D. Microarray technology: an array of opportunities. Nature 2002; 416:885–891.
4. Lockhart DJ, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996; 14:1675–1680.
5. Rosenow C, et al. Prokaryotic RNA preparation methods, useful for high density array analysis: comparison of two approaches. Nucleic Acids Res 2001; 29:e112.
6. Richmond CS, et al. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res 1999; 27:3821–3835.
7. Blattner FR, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997; 277: 1453–1474.
8. Wodicka L, et al. Genome-wide expression monitoring in *Saccharomyces cerevisiae.* Nat Biotechnol 1997; 15:1359–1367.
9. Tao H, et al. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J Bacteriol 1999; 181:6425–6440.
10. Wilson M, et al. Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. Proc Natl Acad Sci USA 1999; 96:12,833–12,838.
11. Diehn M, Relman DA. Comparing functional genomic datasets: lessons from DNA microarray analyses of host-pathogen interactions. Curr Opin Microbiol 2001; 4:95–101.
12. Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. Nat Genet 1999; 21(1 Suppl):48–50.
13. Cohen P, et al. Monitoring cellular responses to *Listeria monocytogenes* with oligonucleotide arrays. J Biol Chem 2000; 275:11,181–11,190.
14. Mandell LA, et al. The battle against emerging antibiotic resistance: should fluoroquinolones be used to treat children? Clin Infect Dis 2002; 35:721–727.
15. Hooper DC. Fluoroquinolone resistance among Gram-positive cocci. Lancet Infect Dis 2002; 2:530–538.
16. Ball CA, et al. Standards for microarray data. Science 2002; 298:539.
17. Brazma A, et al. Minimum Information About a Microarray Experiment (MIAME)—toward standards for microarray data. Nat Genet 2001; 29:365–371.
18. Sherlock G, et al. The Stanford Microarray Database. Nucleic Acids Res 2001; 29:152–155.
19. National Cancer Center (NCI). Center for Cancer Research. Development of Bio-Informatics to

manage access, and analyze cDNA μArray data generated by the NCI/CCR μArray Center. 12/2003. http://nciarray.nci.nih.gov/
20. Cutler DJ, et al. High-throughput variation detection and genotyping using microarrays. Genome Res 2001; 11:1913–1925.
21. Kato-Maeda M, et al. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res 2001; 11:547–554.
22. Hayashi T, et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 2001; 8:11–22.
23. Wong RM, et al. Sample sequencing of a *Salmonella typhimurium* LT2 lambda library: comparison to the *Escherichia coli* K12 genome. FEMS Microbiol Lett 1999; 173:411–423.
24. Himmelreich R, et al. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res 1997; 25:701–712.
25. Tettelin H, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
26. Parkhill J, et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 2000; 404:502–506.
27. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299–304.
28. Hiramatsu K, et al. The emergence and evolution of methicillin-resistant *Staphylococcus aureus*. Trends Microbiol 2001; 9:486–493.
29. Kuroda M, et al. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 2001; 357:1225–1240.
30. Ochman H, Jones IB. Evolutionary dynamics of full genome content in *Escherichia coli*. EMBO J 2000; 19:6637–6643.
31. Hakenbeck R, et al. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. Infect Immun 2001; 69:2477–2486.
32. Tettelin H, et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proc Natl Acad Sci USA 2002; 99:12,391–12,396.
33. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 2001; 98:31–36.
34. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. J Comput Biol 2001; 8:557–569.
35. Ideker T, et al. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol 2000; 7:805–817.
36. Eisen MB, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998; 95:14,863–14,868.
37. Tavazoie S, et al. Systematic determination of genetic network architecture. Nat Genet 1999; 22: 281–285.
38. Ben-Dor Shamir AR, Yakhini Z. Clustering gene expression patterns. J Comput Biol 1999; 6: 281–297.
39. Tamayo P, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999; 96:2907–2912.
40. Yeung KY, et al. Model-based clustering and data transformations for gene expression data. Bioinformatics 2001; 17:977–987.
41. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press, 1995.

42. Tjaden B, et al. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. Nucleic Acid Res 2002; 30:1–7.
43. Wassarman KM, et al. Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev 2001; 15:1637–1651.
44. Selinger DW, et al. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. Genome Res 2003; 13:216–223.
45. Kapranov P, et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science 2002; 296:916–919.
46. Plumbridge J. Control of the expression of the manXYZ operon in *Escherichia coli*: Mlc is a negative regulator of the mannose PTS. Mol Microbiol 1998; 27:369–380.
47. Wassarman KM. Small RNAs in Bacteria. Diverse regulators of gene expression in response to environmental changes. Cell 2002; 109:141–144.

23

微生物种群基因组学与生态学

Edward F. DeLong

引言：历史基础

微生物多样性与生态学

早在35亿年前，微生物就与地球上的生命史和生命功能密不可分。地球的元素循环，包括碳、硫和氮素循环，由于微生物活动而保持平衡和正常功能，有讽刺意味的是，当生态系统正常和地球化学循环功能保持稳态的时候，往往很少注意到微生物的存在，即使注意也是间接的，这就是人们经常忽略微生物对生命进化和对当代地球环境重要性的部分原因。在大多数情况下，人们依旧以人类为中心的角度看待微生物，把注意力集中在相对很少、能导致人类疾病和提供有用产品及工艺潜力的几个物种上。虽然，微生物在地球地质演化、气候变化、生物地球化学及生物进化等方面起核心作用，但也只在近年来才有少数专业人士意识到微生物的广泛重要性。

人们能从许多不同的角度看到古生菌和真细菌的生物多样性，微生物的分类学定义为详细的分类目录提供了背景，在传统上，它基于对基因型和表型的描述。此外，人们可能经常将微生物置于广泛的功能性种群或同资源种群（guild）中（例如硫酸盐还原菌、硝化细菌、甲烷营养菌、光氧营养菌），这种功能性种群或同资源种群的划分，是以微生物在自然环境中的特殊功能特征，或者它在生态学中的作用作为参考，这两种方法都为不同的目的提供了重要信息，但每种生物学定义都有其固有的不足，正如这里所提及的，当代技术，包括分子种系发育和基因组技术，对传统分类学和功能学全面描述微生物自然世界的真正本质提出了疑问。

由于物种概念[1]并不适用于原核生物，因而，对古生菌和真细菌分类学的评价主要基于应用性定义，与后生动物不同，对微生物种群系统的定义、鉴别，搞清相互关系的明显表形描述特征很少，而微生物化石的记载不足，也为进一步了解古代微生物提供广泛而有价值的证据。因此，原核生物分类体系主要或几乎全部依赖表型特征和一些诸如革兰氏染色、鞭毛排列类型、DNA 中 G+C 含量等集群特征（往往相互不一致）。在基于表型特征的分类体系中，没有几个特征能普遍用于分析所有分类类群之间的相互联系，DNA-DNA 杂交已广泛应用于种的鉴定，但仅能用于研究亲缘关系非常近的种或株。其他方法，包括多位点序列分型和 DNA 微阵列技术（参看22章），也成为现代微生物分类学中的重要工具[2]。

更多定位于生态学对微生物界的描述，将微生物分成不同的功能性种群，而这些功能性种群，常与它们的进化或分类学上的联系无关，这种看法可能在经验判断中十分有

用，因为，微生物界包含许多与生物地球化学相关的功能多样性，而这种功能多样性在大型真核生物群中未发现。从微生物生态学家的观点看，功能是主要问题，通常同资源种群以一种有用的方式描述微生物的活动和规则，然而，范围很广的功能性种群，经常将不相关的微生物因其有相似功能而集合在一起，实际上，这些微生物的进化起源和机制特征根本不同，并且，将一种微生物编入某种单一功能性种群的简单想法，忽视了有机体的复杂性，也忽视了它与周围种群及环境的多重相互作用。

直到最近，微生物进化关系和真正的生态学多样性，仍然没有在本质上被人们所了解和描述。要描述天然微生物的生活特征有几个主要障碍，首先，对微生物种的特征描述，在很大程度上依赖于分离纯化技术，尽管这些技术十分成熟，但用这种方法成功获得的仅仅是现存微生物的很小一部分；其次，即使是那些常规培养的微生物，纯培养方法也会给微生物遗传多样性造成大量的取样疏漏。

它们的微观特性给观察带来巨大困难，所以大多数微生物仍然只能通过间接手段观察，观察微生物生态系统的方法通常是将它们破坏使之紊乱，与变幻莫测的海森堡测不准原理类似，在研究微生物世界的过程中，大多需要改变它，缺乏鉴定和描述自然发生微生物的方法，这已是长期存在的问题，并且该领域方法的发展一直是个热点，富有成效解决这些固有困难的方法，就是将分子生物学技术用于研究自然发生的微生物的多样性。

环境中免培养分子种系发生学研究

正如 Zuckerandl 和 Pauling 提倡的[3]，当开始利用信息分子（信息载体）作为进化史证据时，人类在推断所有生命间进化关系的能力发生了剧变。现在，能从所观察直系同源性（orthologous）大分子间的序列差异推断进化关系，对所有细胞形态的生命同源性大分子进行比较首次成为现实。

20 世纪 80 年代早期，比较分子种系发生学的发展，为准确描述自然微生物的多样性扫清了障碍[4,5]，这些免培养法的研究，直接从天然微生物群体的核酸中提取与分子种系发生学有关的信息基因序列。从混合微生物群体中分离纯化 DNA 并进行酶切，在重组克隆中回收，然后对重组克隆进行筛选、分拣和测序。分子序列比较为原始种群组成提供了分子种系发生学的鉴定依据，由于小亚基 rRNA 基因的普遍性和保守性，它们已最广泛用作分子种系发生学的标记物。

自然微生物种群分子种系发生学研究，也为特殊群或种的鉴定提供有用的标记物。作为核酸杂交探针的小亚基 rRNA，已经证明在分子标记方法中非常有用。在使用荧光标记寡核苷酸探针时，相对高水平的胞内 rRNA，为个体细胞分子种系发生学鉴定提供足够的靶标[6,7]。现在，能用着色探针染个体微生物细胞，并能通过荧光分子原位杂交技术进行鉴定，针对不同分类学水平（如界、属、种），现在能够设计嵌套探针，它们可用于天然微生物群体复杂等级分类学的划分[7]。事实上，这种方法已使微生物的分类鉴定与其功能相联系成为可能，现在通过质谱分析，能把分子种系发生学鉴定（通过荧光分子原位杂交）与个体细胞稳态同位素鉴定结合起来[8,9]。

免培养法分子种系发生学研究方法的发展，为环境微生物基因组学的进一步研究奠定了基础。图 1 所示是从自然菌群中克隆大片段 DNA 的一般程序，它是由 Pace 及其同

事在 1985 年[4,5]首先提出，虽然，今天有更新的方法可用，但图 1 所示的流程构成当今微生物种群基因组学研究的基础。最初方案是用 λ 噬菌体作为载体保存自然种群 DNA，在研究海洋细菌浮游生物分子种系发生学多样性早期使用了鸟枪法文库，并证实这种通用方法确实有效[10]。

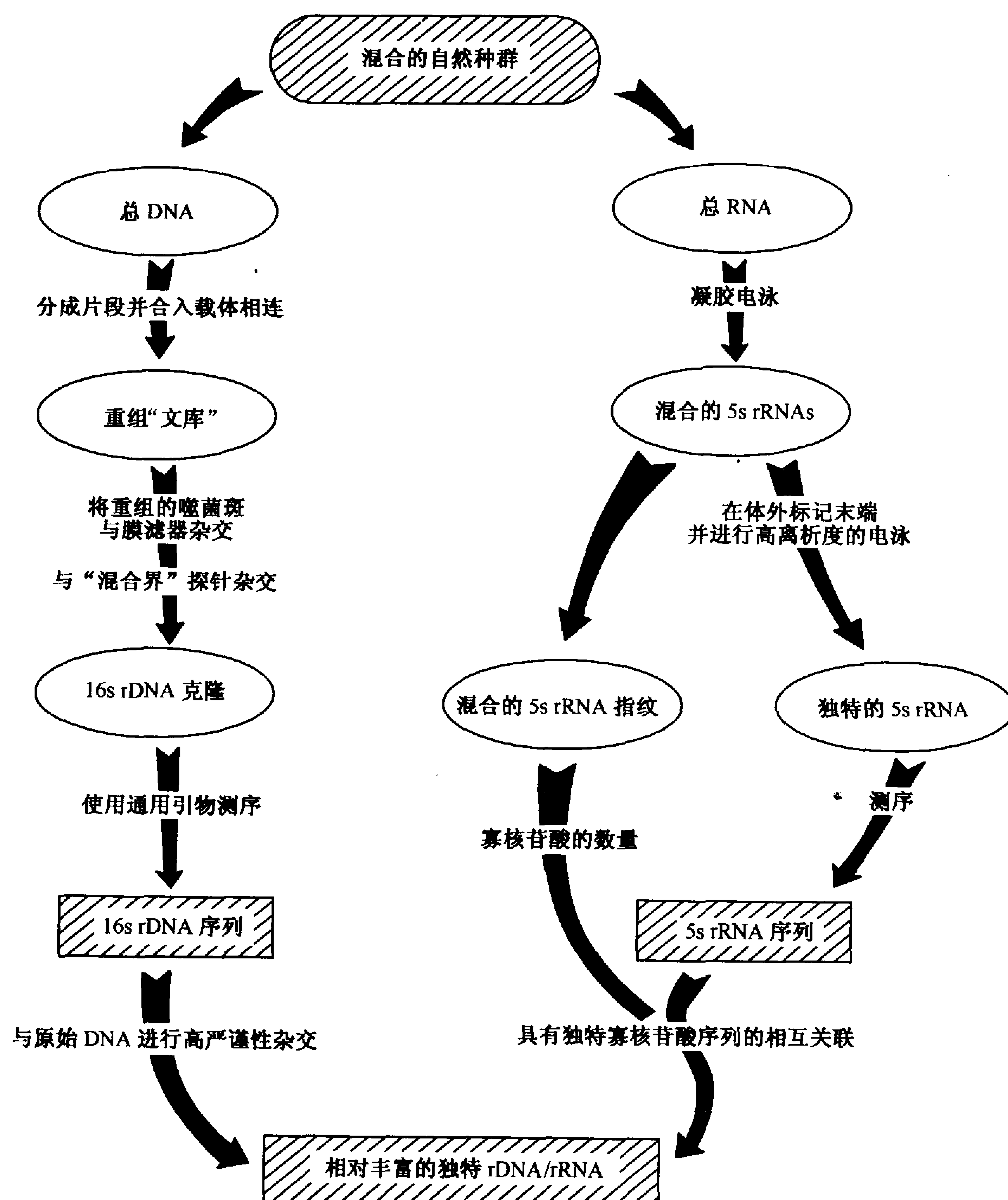


图 1　根据 Pace 及其同事（1985）最初提出确定野生微生物种群种系发生基因调查的方案，包括为种系发生学提供信息的基因免培养环境调查（美国微生物学会授权重印）。

20 世纪 80 年代后期，PCR 技术与热稳定 DNA 聚合酶[11]的联合应用，对微生物生态学领域的发展产生了巨大影响，由于 PCR 技术操作十分简便，因此，该技术成为使用 Pace 程序（图 1）研究自然发生微生物生物多样性的主要工具。PCR 技术真正首次

应用，是从海水混合微生物群系中发现了普遍存在含量丰富的新微生物群体，并证明分子种系发生密切相关类型中存在遗传高度微不均一性（microheterogeneity）[12]。在这点上，大多数微生物生态学家摒弃了操作更加复杂的鸟枪文库法，取而代之以 PCR 扩增技术，从混合微生物群系中扩增单个遗传基因座（如 rRNA 基因），这大大加快了对自然发生微生物分子种系发生多样性的研究，但对描述其功能和特征却无多大的作用。

免培养方法的建立使微生物生态学领域的研究在过去的 20 年中极为活跃。直接从环境中得到的 rRNA 基因的分子种系发生的学比较很快成为研究自然微生物多样性的标准方法[7,13]。免培养方法已经发现了许多新的微生物分类单元，从新的种，到新的门，甚至新的界。这些新近被鉴定的微生物在环境中并不是微不足道的，它们经常代表在陆生和海洋生态系统中，均存在的主要分类单元。这些研究说明，栖息在地球上的最丰富的微生物，其分子发生系统学的特征仍需进一步鉴定。

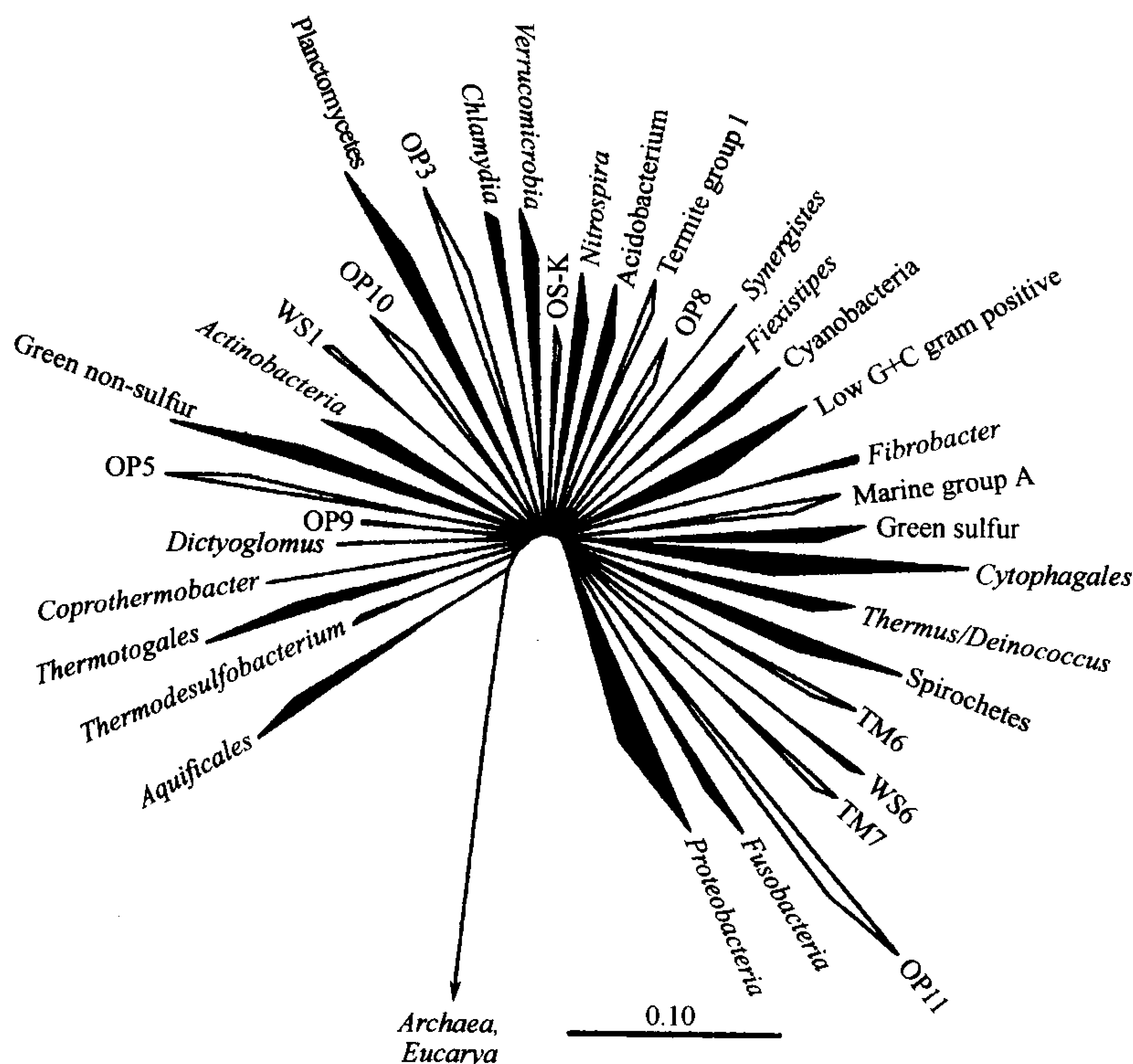


图 2 根据文献［15］提出的主要细菌种系发生学分支。培养典型菌种的细菌分支以实线表示，来自环境样品仅以 rRNA 基因鉴定的细菌分支用空心线表示。标尺代表每个位点的替换频率（美国微生物学会授权重印，见参考文献[15]）。

1987 年，Woese 在他关于细菌进化的论文中，根据当时已知 rRNA 序列数据在细菌界鉴定了 12 个主要种群[14]，这些种群至今仍然代表了大多数已被常规培养并用培养方法描述特征的分类单元。然而，最近免培养分子研究发现，细菌其实有更多的种群分类，但这些种群只有很少或根本没有可供培养的代表菌种（图 2）。目前，现行的细菌

进化树，包括 40 多种在分子种系发生学上分类明确的细菌种群[13,15]，在细菌域新近发现的大多数种群没有可供培养的代表菌种，这表明以前的细菌培养物保藏中心，仅仅展现了一幅关于地球现存微生物的不完全画面。由于分子种系发生学研究，开始为新种培养提供信息与指导，这种情况可能会有所改善[16,17]，至少这些努力会提供主要来自生态学和分子种系发生学一些种群，而目前尚无可供培养代表种的菌株。

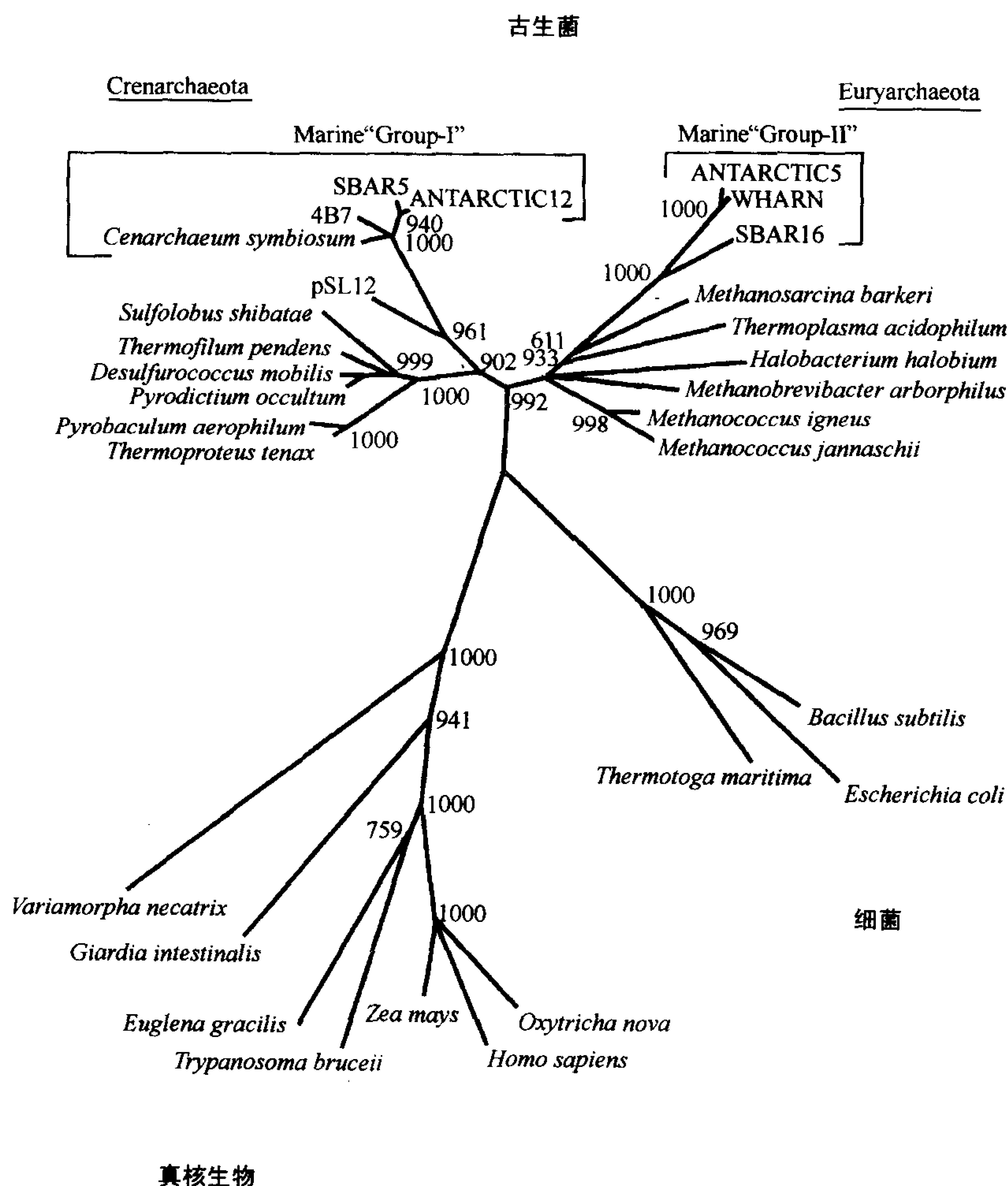


图 3 根据文献［21］展示最常见培养和免培养的海洋浮游古生菌。进化树上展示了两个常见的海洋古生菌群（海洋群 I 和海洋群 II）。（美国微生物学会授权）

免培养法研究已经揭示了古生菌中的新成员，现在，已经知道这些成员广泛而丰富地存在于多种多样的环境中，免培养法研究的一个尤其突出成果，是发现古生菌在非极端环境中广泛存在。以前，人们不相信古生菌与有氧的海洋或陆地环境的生态学存在明显关系，然而，现在很明显，古生菌中至少有两个主要类群（图 3）是海洋浮游生物最

普遍、最丰富的组成[18~22]。在意想不到的环境中，包括陆地和海洋，依然不断发现新种群[18,19,23,24]，已经清楚地知道，在海水深度 100～5000 米范围内，古生菌通常以 1×10^5/ml 密度存在[20,22]，以这样的细胞密度，古生菌代表约 20%或更多海洋微生物，而 10 年前人们认为，那里根本不存在古生菌。让人预料不到古生菌多样性的最近例子，是纳古生菌（*Nanoarchaeum*）[25]的发现和培养，它专性寄生于 *Igneococcus* 属的古生菌中，新近发现的这一古生菌与已知种群差别很大，它的 rRNA 序列不能经过通用寡脱氧核苷酸引物进行测定。

环境中丰富的微生物新种鉴定，并不是免培养法研究促成的惟一发现，免培养法 rRNA 研究，在亲缘关系非常接近的微生物种群内和种群间，也展示了数量巨大的微多样性（microdiversity）和微不均一性[12,26~28]。这种遗传的微不均一性（在单一种群中观察到高度相似但又不同的 rRNA 序列）与特异分类群相对应的环境或所用方法无关。这种 rRNA 的微变异（microvariation），怎样与高度有序的基因组结构和组装相对应，这样的变异怎样对微生物种群的动态产生影响，以及哪种机制产生并维持这种变异，都值得进一步探讨。

基因组学与微生物生态学

把基因组学和微生物生态学这两个完全不同的学科联系在一起，有许多概念上的障碍，基因组学由技术推动，在近 5 年内发展成了一门学科，它的优势在于，不用筛选就可以直接处理大量数据。相反，微生物生态学已有数百年历史，但它仍然是一门受技术限制的学科。基因组学的技术优势与微生物生态学所提出的众多问题相结合，可使二者相互补充，相互推动。有时，侧重于实验室和模型系统的基因组学，与基于实地的微生物生态学相接触时会出现一些摩擦。

虽然，这两个学科的界面并非总是一致，但很明显，它们有很多地方相互补充。微生物生态学家可以从基因组学的高通量信息处理中受益，相应地，微生物基因组学也可从适当环境条件下，基因组进化动力学的直接研究中受益（经过与微生物生态学家共同努力）。

当代基因组研究法

正如本书第 1 章所述，自动化 DNA 测序技术发展到一个反应能读 500bp 时，用鸟枪法测序全基因组成为可能，这种表观流通量可使鸟枪法测序经 24304 次序列阅读，就可以完成整个微生物（流感嗜血杆菌的）的基因组测序[29,30]。从那时起，比原来处理量增加了很多倍，更加快速，廉价和容易掌握的 DNA 测序技术不久就产生了，于此前后，序列集成软件和自动化解码技术也在发展，当然，可能不一定与测序技术同步发展。

能够稳定地保存和复制大片段 DNA 新载体的发展，是基因组革命中的另一个里程碑，尤其是 20 世纪 90 年代初发展的，在人类基因组计划中应用的细菌人工染色体组（BAC）载体特别有用，细菌人工染色体组中的 DNA 复制，受大肠杆菌 F 质粒的 F1 复制起始位点控制[31,32]，由于拷贝量低，BCA 可稳定复制大片段 DNA，而在其他载体中

这些大片段不稳定[32]。

到目前为止，大部分（大约 75%）已测序基因组属临床上的重要细菌，在所有原核生物序列中，大约 90%是细菌。古生菌基因组序列在数据库中很少出现，正如本书中通篇详细论述的那样，从完整原核基因组序列可学习到如下内容：

(1) 发现大量编码未知功能的新可读框。

(2) 对水平基因转移在基因组（宏观）进化过程中重要性的认识。

(3) 与致病性有关基因组特征的鉴定。

(4) 由于专性共生或寄生而造成基因组序列的分化。

这些的确是很重要的规律，但事实上据估计，仍有 99.9%以上的野生微生物是未能人工培养、未知或正在研究中。要想了解肉眼看不见的微生物世界，需要采取与过去完全不同的方法，微生物种群基因组学是其中的方法之一。

基因组研究方法在自然界微生物种群中的应用

目前，一部分微生物全基因组序列来自未经纯培养的微生物，这些微生物包括病原体，如麻风分枝杆菌（*Mycobacterium leprae*）、普氏立克次氏体（*Rickettsia prowazeki-i*）和恶性疟原虫（*Plasmodium falciparum*）（见第 20 章），这些病原体可以从其他受感染细胞中纯化出来，但必须在含有动物寄主细胞的条件下进行寄生培养。另外，两个布氏蚜虫（*Buchnera aphidicola*）中的蚜虫内共生体菌株的全基因组序列已见报道，在这个例子中，布氏蚜虫内共生体 DNA 含量很高（每个细胞含 100 个基因组拷贝），因此，在进行直接全基因组鸟枪法测序前，从 2000 个蚜虫的含菌胞（bacteriocytes，含内共生体的昆虫器官）中提取内共生体，并用 5μm 滤器对释放的内共生体纯化[38]。

上述应用证明，现代基因组测序方法，如全基因组鸟枪法克隆和测序，完全可以用到未经纯培养的微生物，到迄今为止的很多例子中，在鸟枪法测序前，必须将靶细胞从其他组织中进行物理纯化，随后的全基因组测序方法和实验室培养菌种所用的方法相同。

接着有个很明显的问题，相似方法和技术可用于自然界中混合微生物种群吗？在理论上，用于单一菌种基因组分析的方法、技术和分析策略，完全可以用于天然微生物种群中。在进行野生微生物基因组鉴定前，要搞清楚以下几个问题：从环境中取得天然微生物全基因组序列，特殊靶细胞的物理纯化与分离是必需的吗？大片段基因克隆与测序技术，能用于自然界微生物种群的免培养策略吗？既然天然微生物种群中广泛存在遗传微多样性，现有技术能从鸟枪法测出的复杂混合种群微生物 DNA 中，整理出可信的全基因组序列吗？

环境微生物基因组学：方法学

样品采集

采样方法总是依赖于整个方法的选择，并由采样环境和被取样微生物种群的特性及与它所处的环境决定，设计采样方法时，要考虑很多问题。

微生物细胞需要从土壤、沉淀物和岩石矿脉等样品中分离纯化吗？在 DNA 纯化和下游进行酶修饰的步骤中，污染物的干扰是考虑的主要问题。在进行 DNA 抽提前，微

生物细胞需要以某种方式进行富集吗？如有些操作步骤，最少需要 1×10^9或更多的细胞量，这主要取决于所研究的微生物种群。采用多大的 DNA 片段和什么样的克隆方法？这主要看细胞是否需要嵌合到琼脂糖中，以保持高分子质量和完整的染色体组（如果插入片段为 40kb 或更小，则没有必要将细胞进行琼脂糖包埋）。所测样品的复杂性、丰度与均匀度和特殊组成成分如何？例如，少量真核生物基因组，就可影响原核生物基因组序列在任何给定文库中的比例，因为前者的基因组大小是后者的 10～50 倍。

在进行环境基因组学研究之前，所有这些问题及许多其他问题都是非常重要和值得考虑的，对原始样品特性、复杂性和组成成分的了解，包括基质和微生物，将在很大程度上影响实验设计和文库构建以及下游分析的最终结果。

文库构建方法

本书中深入讨论的鸟枪测序方法（见第 1 章），到现在才被复杂的自然界微生物种群测序所重现，到目前为止，鸟枪法测序在微生物种群分析中，未广泛应用的主要原因是，为达到一定覆盖率，种群测序规模（和经济因素）是不确定的。虽然如此，鸟枪法测序将毫无疑问地在微生物种群基因组学中扮演重要角色。目前，只报道使用 λ 文库、粘粒文库或 BAC 文库等方法，下面简要讨论插入大片段 DNA 克隆方法在自然界微生物种群中的应用。

一旦 λ 抽提物经包装转染到大肠杆菌后，Fosmids 在序列和其他特性上与 BAC 在本质上一致，BACs 和 Fosmids 的主要区别是二者导入大肠杆菌中的方法不同，BACs 经电转化到大肠杆菌，而 Fosmids 先包装进入噬菌体头部，然后再转染到大肠杆菌寄主细胞。这些方法上的差异影响装载重组 DNA 插入片段的大小，BAC 从自然界微生物种群装载片段大小可达 200kb，而 Fosmid 装载 DNA 插入片段大小在 32～45kb 之间。

BAC 克隆方法见图 4，该法与 20 世纪 90 年代初所介绍的方法相同，简而言之，为防止下游细胞裂解和纯化步骤过程中的物理剪切作用，需要将生物细胞包埋进琼脂糖中。尽管这种方法十分有效，但其他与细胞共同纯化的物质也能包埋进琼脂糖中，并且有可能抑制下游处理过程中的酶修饰作用（例如，抑制限制性内切核酸酶的酶切和酶连)。因此，进行琼脂糖包埋的细胞要尽量少含污染物，从干净水体中富集细胞，比那些污染物较多的环境，如土壤和沉积物中得到的细胞更适合包埋。

用琼脂糖包埋细胞进行蛋白酶 k 去污剂消化后，释放出的 DNA 通常用 *Bam*HI、*Hin*dIII 等限制酶部分酶解，产生大量适当大小的 DNA 片段，其范围在 100～300kb 之间。通常，在开始准备过程中会产生剪切的 DNA 片段，因此，要小心控制部分消化条件，使部分消化产生的 DNA 片段大小在适合范围内（见图 5A)，纯化（经脉冲电泳）后从琼脂糖中回收 DNA 片段，连接到 BAC 载体，然后通过电转化导入大肠杆菌。

筛选技术的发展和电转化条件的优化提高了整个过程的效率，最重要的是，在这一普通操作规程中，对任何不同检测样品都要进行优化，例如标准 BAC 文库，保证足够微生物细胞数量很重要，在最初的琼脂糖填料中一般需要最低浓度为 1×10^{10}个细胞/毫升。

虽然，装载片段平均长度较小（40kb)，但从环境样品中构建高质量文库时，与标

准的 BAC 克隆方法相比，Fosmid 克隆方法有几个主要优点，它所需要的 DNA（平均片段小）量少，在分离大量完整染色体组中大大降低了对细胞（1×10^{10}个）（如本节所讨论的）的大量需要。最初的 Fosmid 克隆步骤见图 5A，在该方法中，载体臂和用 *Sau*3A 部分消化的 DNA 都已准备好，一般还要进行片段大小的筛选，经过酶连、λ 抽提、包埋和转染，文库保存在微量滴定板中，可用质粒纯化和筛选的标准步骤进行。

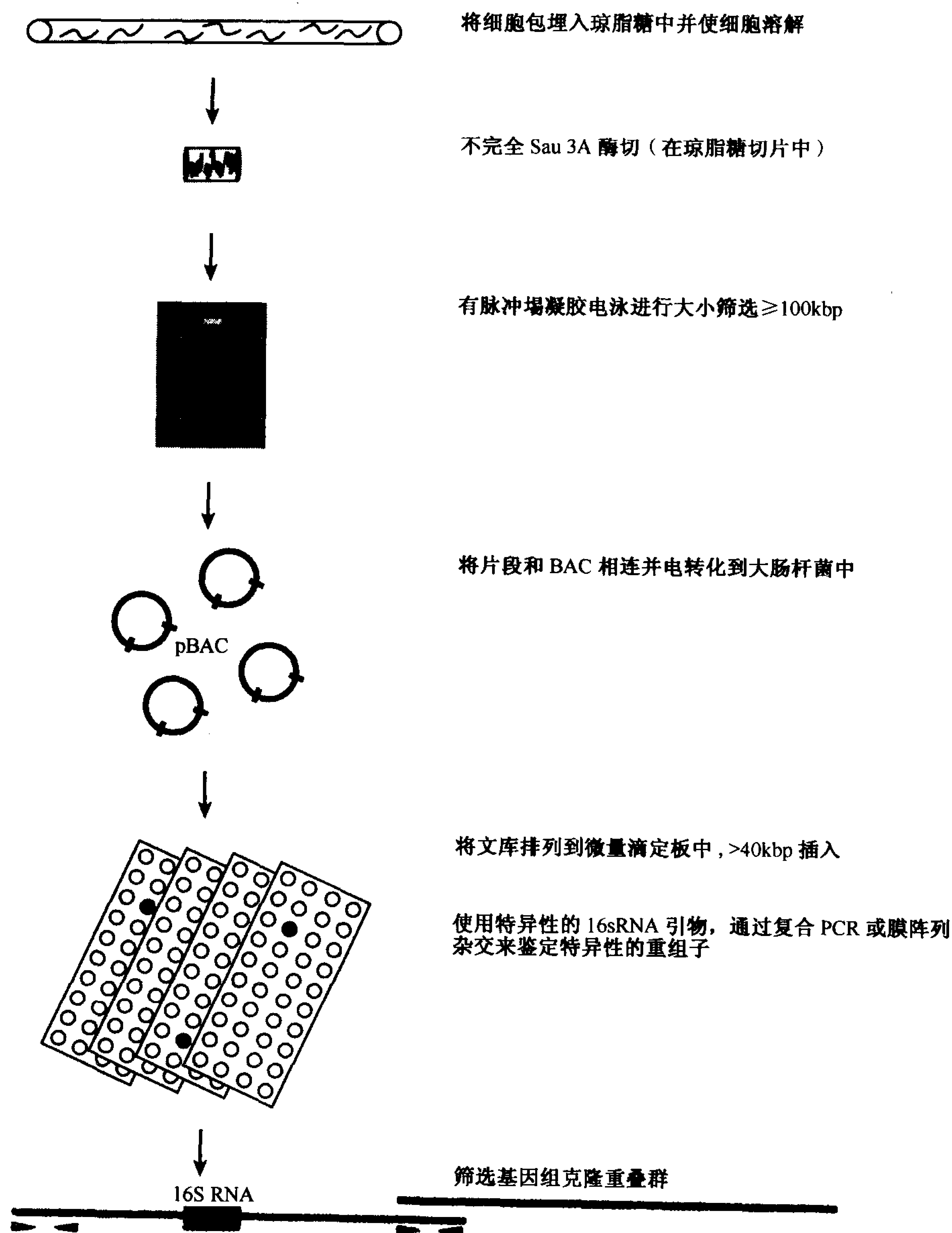


图 4　细菌人工染色体文库构建流程（PFGE，脉冲场凝胶电泳）。

新方法大大提高了 Fosmid 文库的构建效率，这些新方法包括，DNA 插入片段的随机物理剪切、末端修饰和环化 Fosmid 的平末端连接，这与 BAC 克隆的部分限制性内切

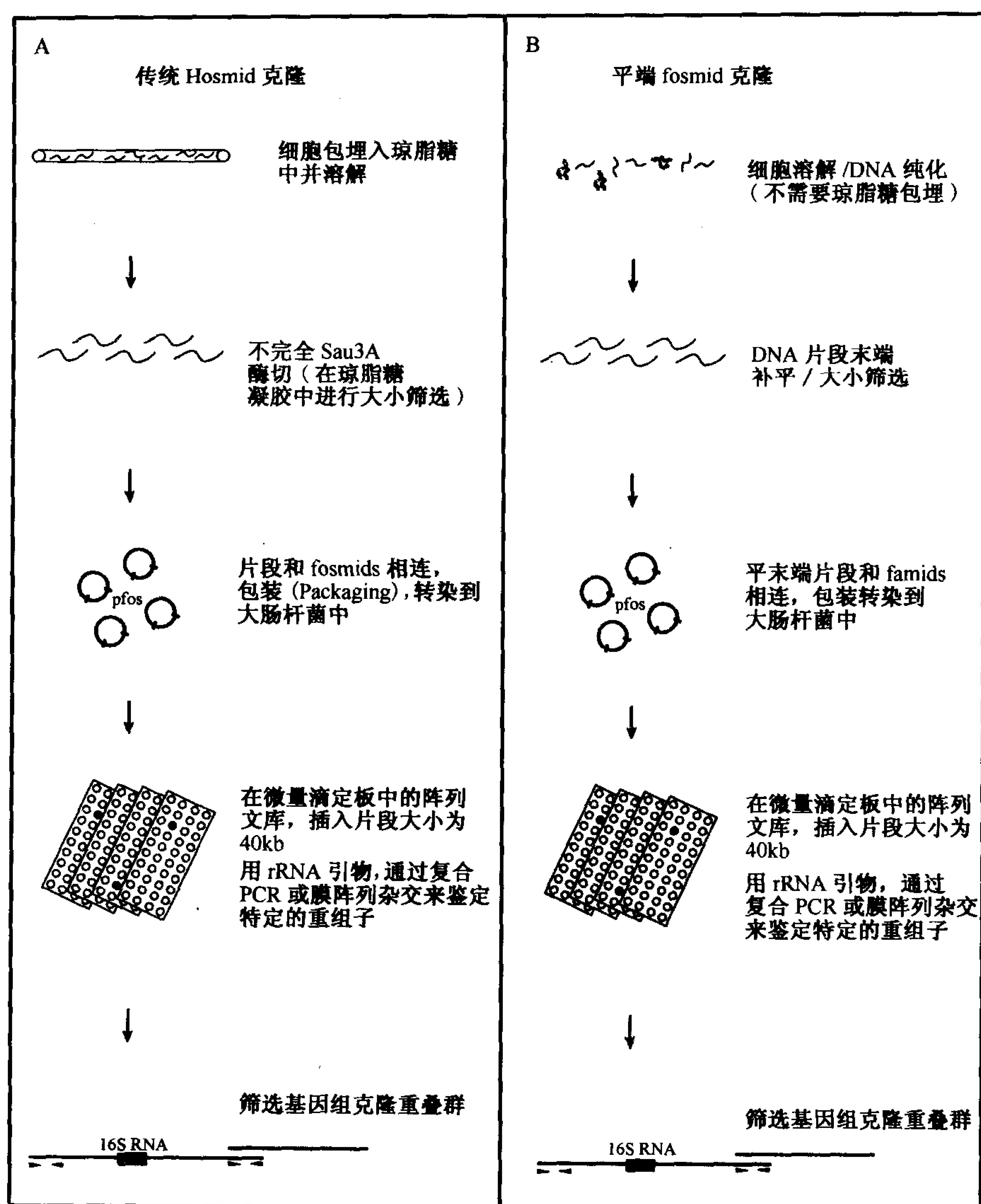


图5 Fosmid 文库构建流程。(A) 传统的用部分降解 DNA 和黏端克隆法构建的 Fosmid 克隆。(B) 平端 Fosmid 克隆策略。

核酸酶消化和标准 BACs 载体的黏性末端克隆不同（图 4 和 5A）。一般情况下，标准纯化操作规程能产生所需大小的 DNA 片段（40kb），可省去细胞的琼脂糖包埋，部分限制性内切核酸酶消化和脉冲场凝胶电泳等步骤（图 5B）。此外，更加严格的 DNA 纯化方法（如 CsCl 平衡密度梯度离心），可成功地用于平末端 Fosmid 克隆方法。

Fosmid 克隆方法有很多优点，可用较少样品构建高质量 fosmid 文库，甚至，平末端 fosmid 文库比标准 BAC 文库更具有代表性，因为有较少 *Hin*dIII 和 *Bam*HI 位点的片段，在部分消化法准备的 BAC 和 fosmids 中表现不佳，而平末端 fosmid 克隆方法则不是这样。按照这种流程，准备 1μg 或更少的自然界种群微生物 DNA 构建高质量的 fos-

mid 文库完全可行。

在文库构建中，应着重注意存在不同生物 DNA 组成嵌合 BACs 的可能性，虽然这是个潜在问题，但在文库构建中所采用的方法，特别是克隆前 DNA 片段大小的选择，可减小已知风险，另外，用 λ 包装将 fosmids 导入大肠杆菌的方法，包含有片段大小选择的步骤，这进一步减小了含嵌合 BAC 的风险。

另外几种方法都可排除 BAC 是嵌合体的可能性，BAC 种群筛选法可鉴定文库中同源但不等同的 BAC，这代表给定文库中关系密切但并非相同的菌株[42,43]。这些 BAC 亚种含有很相近的染色体结构，包括基因组成、组织形式和共线性，事实证明，上述基因组结构并不是人为造成的。另外，高分辨率光学绘图方法[44]正在应用，它可以测验和证明单个克隆子的完整性，与之相似，基因组绘制技术[45]也可用于鉴定单个克隆子的完整性和双末端 BAC 序列的存在。

文库筛选和分析

以前发展针对普通质粒的标准筛选方法，对 BAC 和 fosmid 都适用。高密度菌落印迹法（宏阵列）是一种典型的方法，在 20cm×20cm 膜上，可筛选 10 000 个克隆子[40]，PCR 多重筛选是另一种筛选特定重组克隆的快速有效方法，这两种方法都成功用于鉴定环境 BAC 和 fosmid 文库中目的克隆[39,40]。为了筛选 rRNA 基因，应用其他快速多种形式的筛选方法：自动化毛细管电泳，包括长度异质多态性[46]或末端限制性片段 PCR[47]筛选。

此外，功能基因筛选在获得具体代谢途径和过程信息方面十分有用[48]，已证明非常有用的另一筛选方法是随机末端测序，它可以用作文库质量的评估、基因调查、重叠群鉴定和种群比较分析（例如 http: www.tigr.org/tdb/MBMO/BAC-end-ann-into.shtml）。

虽然，单拷贝 BACs 和 fosmids 的末端测序曾经极具挑战性，但随着测序技术的发展，该技术上的困难已被克服，达到了高通量 BAC 末端测序，为获得任一个既定环境文库质量和数量方面的评价，这也许是首要选择的方法。

文库评估中非常重要的是，甚至在低拷贝的 BAC 载体中[49]，也可能存在不同基因组序列的不同覆盖率。造成这些假象的原因有：存在高度重复的串联序列元[49]，或因为重组 BAC DNA（尤其细菌 DNA）在大肠杆菌中的表达[50]，有时这是致命的弱点。在很多情况下对从复杂样品中获得基因组的保真度进行评估很重要，但要精确的量化仍有很多困难。

通过 rRNA 基因筛选文库内种系发生的代表基因，是评价文库中基因组多样性的有用方法[51,52]，用统计方法估计种群中基因组的回收情况或许有用[53]，但可能需要大量样品。

重组文库中基因组的回收程度，很大程度依赖于研究目的，对于基因组种群生态研究，样品的归一化是不必要的，因为代表性基因组的数量信息是主要的；相反，进行生物勘探的研究者（bioprospector）总希望获得基因组的最大多样性，因此，想方设法扩大文库中低丰度基因组类型的数量。

微生物群体基因组学和生态学在真实世界中的应用

在各种动机推动下，研究人员致力于微生物群体基因组学（也称为环境基因组学(environmental genomic)[40,42]，或泛基因组学（metagenomic[51]）的研究，在好奇心驱使下，生态或群体遗传学研究、生物勘探以及生物地球化学或生物地理学研究都受益于常规研究方法，微生物种群基因组学研究方兴未艾，并伴随着基因组学研究技术的发展而协调发展。

图 6 是从基因组学到自然微生物群研究的一些应用，基因发现、代谢途径特征及其开发利用、生物化学研究，以及基因型和生态分布之间相互关系，图 6 展示研究流程中的几个方面。下面列举环境基因组学的一些例子，阐明了多种研究方法的潜力，以及它们与其他领域和学科的协同作用，这些研究事例结合起来就完成了图 6 所示的整个流程，并且已经取得了意想不到令人惊奇的结果[39,52,54]。

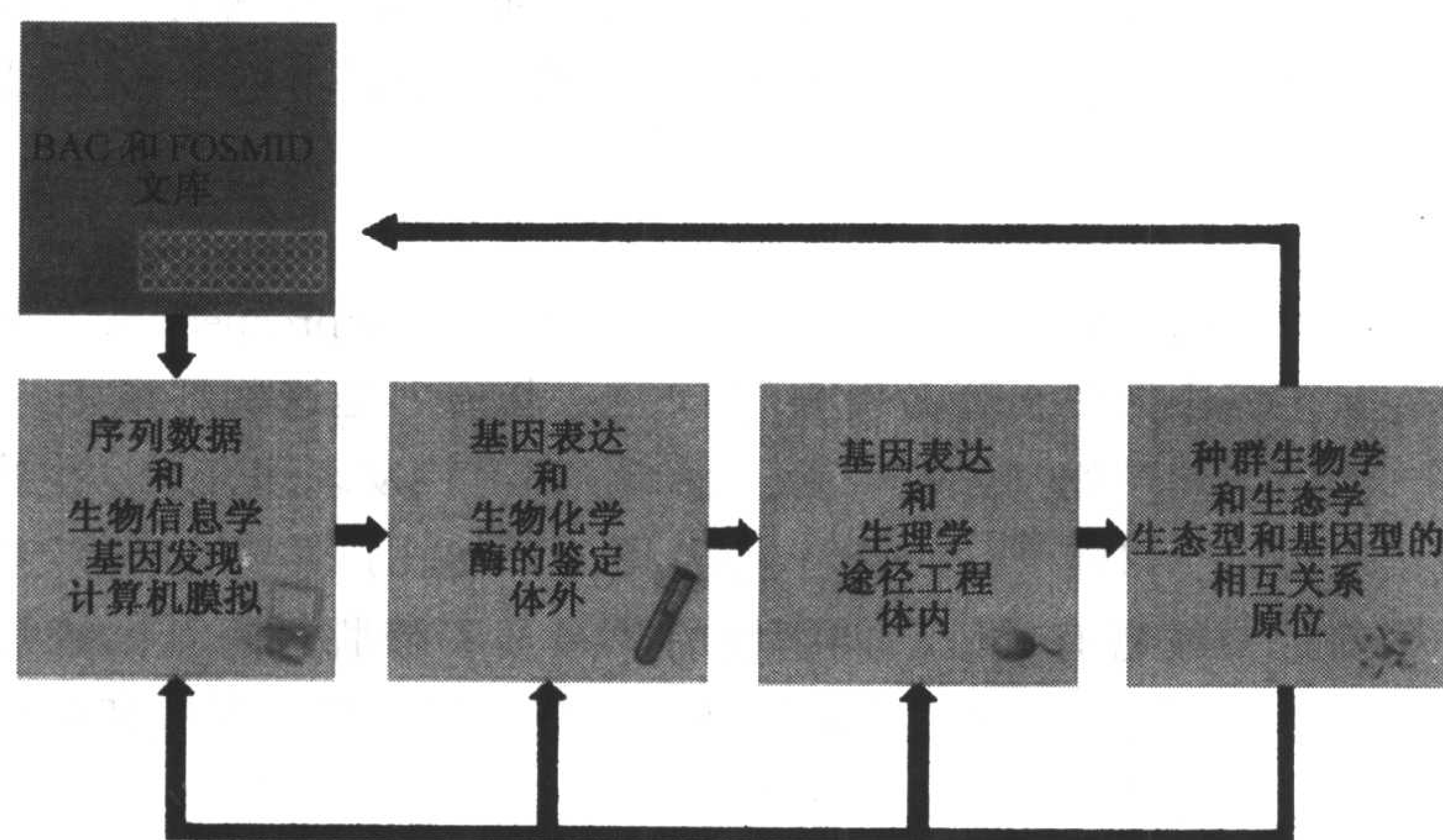


图 6 微生物群体基因组学整体研究流程概括。

定向研究天然微生物的特征

如前所述，对从环境中获得 DNA 进行大基因组片段克隆的方法之一，是采用高保守性和低拷贝数的细菌人工染色体（BAC)[31,32]，1996 年首次报道，利用细菌人工染色体载体，从混合微生物群体中回收 DNA 的研究[40]。在这项研究中，构建的细菌染色体文库来自太平洋 200 米深处的微生物，该文库用于研究未能培养的海洋古生菌，用 BAC 技术从浮游古生菌中克隆了 40kb 的 DNA 片段，并部分测序，这一研究第一次使人们认识到海洋古生菌的基因排列、基因数量和编码蛋白质基因[40]，后来，在同样的 BAC 文库中，发现了浮霉状菌目（Planctomycetales）细菌的大基因组片段[55]。

此后，用相似方法从未能培养的非嗜热古生菌海绵共生体（*Cenarchaeum symbiosum*）中，较系统回收了其基因组片段[42,56,57]。含 rRNA 操纵子的基因组片段也发现可编码古生菌的 B 型 DNA 聚合酶，该酶与嗜热菌 *Crenarchaea* 有较高的序列同源性。*C. symbiosum* 的 DNA 聚合酶已经表达，并研究了其生化特征[56]，其生化催化活性与已研

究嗜热古生菌的聚合酶相同。然而，*C. symbiosum* 聚合酶的热稳定性只能到 40℃，而相应嗜热古生菌的聚合酶至少在 75℃ 下稳定。这表明通过从种系发生标记进行基因步查（gene walking）的方法，对未能培养的微生物基因产物进行鉴别、表达及描绘的应用性[56]。在一项后续研究中，利用 BAC 文库对 *C. symbiosum* 基因组异质性的本质及程度进行了详细研究[42]。

为了评价天然浮游微生物（它们许多都是很难培养）的功能及自然特性，Beja 等[52]利用蒙特利海湾（Montery Bay）表层水中的微生物构建了一个 BAC 文库，这个 BAC 文库的插入序列平均大小为 80kb，一些克隆超过了 150kb，这些结果是一种最大的鼓舞，它们同时表明，从混合微生物群体 DNA 中产生较长 BAC 文库插入序列的可行性。

表层海水 BAC 文库分析发现，种系发生数据与 PCR 扩增 rRNA 克隆文库的数据吻合[52]。一个来自未能培养海洋古生菌的 60kb 基因组片段，在该文库中定位、全序列测定及基因注释（annotated），该文库的筛选和分析表明，利用这种方法可获得大量基因组方面的信息，并可为深入研究天然微生物的多样性及生物特性提供一种手段。

生物勘探

随着众多研究对药物、酶和其他天然产物生物勘探（bioprospecting）的关注，已将大量研究集中在从土壤中回收土著微生物的 DNA，其原因是这些环境中约 99% 的微生物未能培养过，大量天然产物（如抗生素、酶等）还有待发现和开发。生物勘探方法在生物技术领域已应用了十多年[58]，对插入小片段（5～8kb）环境 DNA 文库进行定向生物勘探，在一些前沿领域富有成效，如脂解的 4-羟基丁酸脱氢酶基因或几丁质酶基因的分离[59~62]。采用将环境发现和体外定向进化相结合的方法，寻找和优化工业上重要的生物催化剂（如 α 淀粉酶）也已有报道[63]。

在另外的报道中[51]，直接从土壤中提取微生物 DNA，并克隆到 BAC 文库中，获得两个不同泛基因组文库，并对其加以分析，其中一个 DNA 插入序列平均大小为 27kb，另一个为 44.5kb，该文库中 rRNA 基因的种系发生组成与土壤中的种系发生调查结果一致。尽管这些文库中 DNA 插入序列平均大小并不比传统的 λ 或黏粒（cosmid）文库大多少，但这是对从土壤微生物群体中回收 DNA 的首次报道[51]，并表明其中一小部分 BAC 克隆能表达可确认的一些表现型，包括 DNA 酶、酯酶和淀粉酶的活性，这与以前报道的结果一致[50]。另一项研究致力于从 BAC 文库中获取土壤微生物基因组，所使用的插入序列平均大小为 37kb，范围在 5～120kb 之间[64]，Brady 等类似的报道是从土壤 DNA 黏粒文库中，得到广谱抗生素紫色杆菌素（violacein）的生物合成基因簇[65]。

微生物群体基因组学

环境微生物基因组学促进了微生物群体遗传学的深入研究，并将它们置于生态关系中，现在的研究集中在生境多样化的表现型（即生态型）培养菌株上[26]，或是从环境中定向克隆大片段 DNA[42,43]。将比较基因组学和微生物群体生物学相结合的事例，是最近关于两种相近原绿球藻（*Prochlorococcus*）的比较[66]，原绿球藻是含叶绿素 b 的海洋蓝细菌，它们占海洋中营光合作用生物量的 50%。高光和低光原绿球藻的差异在于，

叶绿素 b 与叶绿素 a 的比率、最佳吸光范围以及它们在水中的相对分布，但它们的小亚单位 rRNA 只有 3%差异。

高光适应型（MED4）和低光适应型（MFT9313）的两菌株进行了全基因组测序[66]，比较分析结果揭示，这两个相近的原绿球藻生态型具有生理和生态差异的基因组起源。低光适应型（2.4Mb）比高光适应型（1.7Mb）的基因组较大，并且具有更多与光合作用装置有关的基因，包括藻红蛋白生物合成基因；相反，高光适应型有更多编码高光诱导蛋白的基因，以及用于紫外线损伤修复的基因[66]，它们的氮同化基因差异也影响这两个菌株在水体中的分布[67]。

群体遗传学还表现在另一个突出领域，在这一领域中，对自然微生物基因组细微特征（和它们微小的变化）的研究非常重要。过去关于微生物群体遗传学的大量信息，主要来源于两种主要的数据类型：多位点酶电泳和多位点序列分型，而且主要集中在病原体[68]，这些研究主要涉及对可培养病原体菌株的比较，而不是真正的群体内或种群间的比较。总之，在这些研究中，培养对取样偏差的影响还不清楚，正如群体遗传学家所承认的那样：取样偏差、生态亚结构以及出现的暂时适应性克隆，都需要认真加以考虑[68~70]。基因组学将以其独特的形式走进微生物群体生物学——采集真实的群体，消除培养所带来的取样偏差。

还有些重要问题，如关于重组和点突变在微生物多样性中贡献的争论，需要微生物群体基因组学予以直接阐述。例如，一项研究表明，在一种单一海洋古生菌群内，虽然其 rRNA 序列分析并无差异，但是，它们都存在大量的基因组变异[43]，具有相同 rRNA 序列的古生菌，在 rRNA 侧翼编码基因上也表现出很大的差异，即使它们共存于同一群体中，这说明自由生活在同一地域的微生物物种中，存在大量的等位基因变异，这是点突变和基因飘移的直接结果。

微生物群体基因组学较好地阐明了水平基因转移、重组和基因飘移在微生物种内和种间多样性的重要性，它能更优化地估计群体遗传学的中心参数[71]，这种估计对理论家很有帮助，也能为我们提供迫切需要和数据可靠的广义模型。这样的研究将会促进新理论的发展，而且提高在不同生物体、生态和群体条件下对微生物群体生物学的认识。

自然微生物的代谢重建

在对蒙特利海湾表层水 BAC 文库的后续研究中（在前文中讨论过），从未能培养的 SAR86 中获得 130kb 的 BAC 克隆序列已经测定，它的 rRNA 操纵子下游序列有一个意想不到的类似视紫质蛋白（图 7[39]），这类膜蛋白以前在细菌中或海洋微生物中从未发现过，当伴随发色团表达时，这种新型膜蛋白称为 proteorhodopsin 证实是光驱动的质子泵[39]。这种蛋白的生化功能和光循环特性，与常见细菌视紫质相似，后者最初是在极端嗜盐古生菌中被发现。这些数据表明，SAR86 浮游细菌是一类遍布全球水域的新型光养生物，SAR86 的光驱动产能机制如图 7 所示，而且，现在假定这些微生物是光能异养型，它通过分解有机碳源获得相当一部分能量[39]，后来的研究表明，这种新型视紫红质有多种遗传变异类型，以适应不同海水深度吸收不同波段的光[54]。

总之，这项研究表明，用自然微生物群体基因组学研究不只是生物信息学的一种尝试，这导致发现和证实了自然生态系统中的新过程和属性。基因组研究为深入细致地研

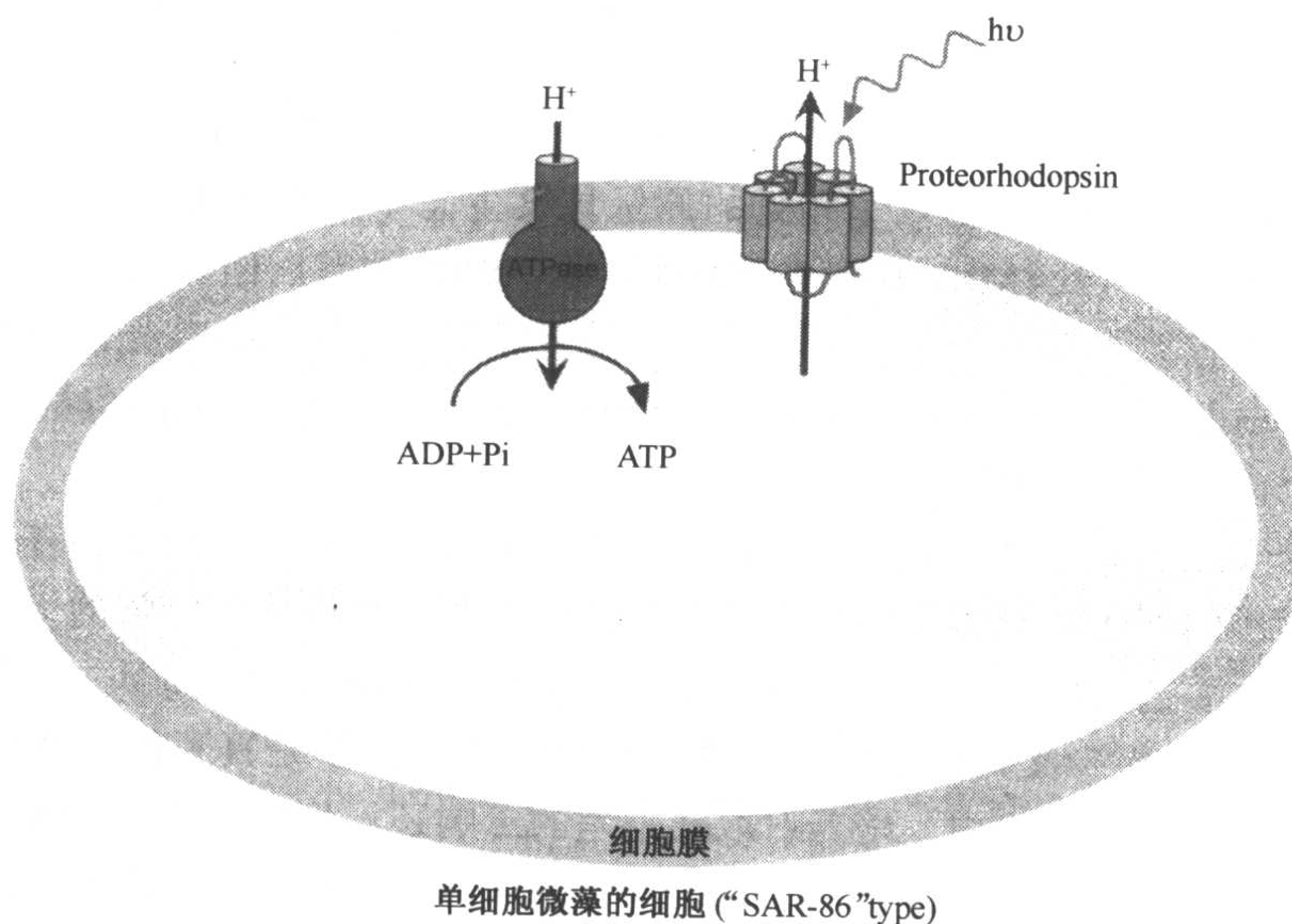


图 7　推测的免培养海洋浮游细菌 SAR86 的能量产生机制。ADP，5′二磷酸腺苷；ATP，三磷酸腺苷；Pi，无机磷酸（引自参考文献［39］以及 E.F.DeLong 和 O.Beja 的未发表结果）。

究新发现的过程提供了工具，这些在图 6 流程图中得到了体现，强调了序列分析、生物物理、生物化学以及生态研究之间的反馈及相互作用。

从蒙特利海湾表层水的 BAC 文库中筛选出含菌绿素的好氧性不产氧的光营养（aerobic anoxygenic phototrophic，AAP）细菌[48]。最近报道，在海水中存在数量惊人具有菌绿素光合系统的细菌[72,73]，表层水中一些微生物的 BAC 克隆含有 40kb 超操纵子，它编码光合作用反应中心、类胡萝卜素以及菌绿素的生物合成 AAP 细菌基因。对 AAP 细菌光合超操纵子的基因组结构分析表明，一些大量存在的光营养浮游微生物，不是现有培养菌株预测的那种类型，这再次表明，免培养方法在认识自然界中微生物的重要性，基因组学方法的应用能在自然微生物群体中有大量新发现。

比较生态系统基因组学

比较生态系统基因组学（comparative ecosystem genomics），在微生物生物学中是一个新兴的几乎尚未开拓的前沿，它将不可避免地从微生物群体基因组学中显现出来。不久的将来，采集一套代表全部微生物群落和生态系统的基因，以及基因组功能和种系发生的样品将成为可能，当基因含量类型与生态系统内或生态系统之间的差异有相关性时，这种比较就很有趣。通过比较，微生物之间极其重要的相互关系非常明显，通过对基因组的仔细研究，将得到有机体相互作用的特异性表征，更进一步检验微小的基因组变异和环境差异之间的相互关系，将使基因组变异的生态学意义以及相应功能的改变显得更加清楚。

另外，对特殊生境中的基因组进行检验，可能发现更高级营养关系以及生态系统特征，简单的例子如图 8 所示，若要进一步了解，请参考正在进行的对蒙特利海湾不同水

域 BAC 文库的分析，包括随机的 BAC 末端序列（图 8；http://www.tigr.org/tdb/MB-MO/）。

现在已经研究清楚，微生物群落在海水水体中成层分布[74]，随着基因组数据的累积，更详尽的认识构成不同群落基因组的本质将成为可能，在此，应考虑的问题包括：在每一深度水域发现的功能基因，是否反映对这些系统生态关系的认识？基因和基因组分布提供有关过程的信息，与对每一生境生态过程的认识相符？不同生境和群落中共享的一套核心基因（图 8 中的交叉区域）有哪些？在各个生态系统中，是否存在虽不同源但却有相似功能的核心生物化学反应和代谢途径？或者说，界定每一深度对应微生物生态系统的基因组特征（图 8 中的白色区）是什么？如何从功能上理解这些生态系统？

样品来源	细胞总数	平时插入片段大小	估算的覆盖大小	累积的基因组
海水 0m	2.5×10^{11}	80 kb	458 Mb	~152
海水 80m	1.5×10^{11}	74 kb	1184 Mb	~390
海水 750m	5×10^{10}	60 kb	96 Mb	~32

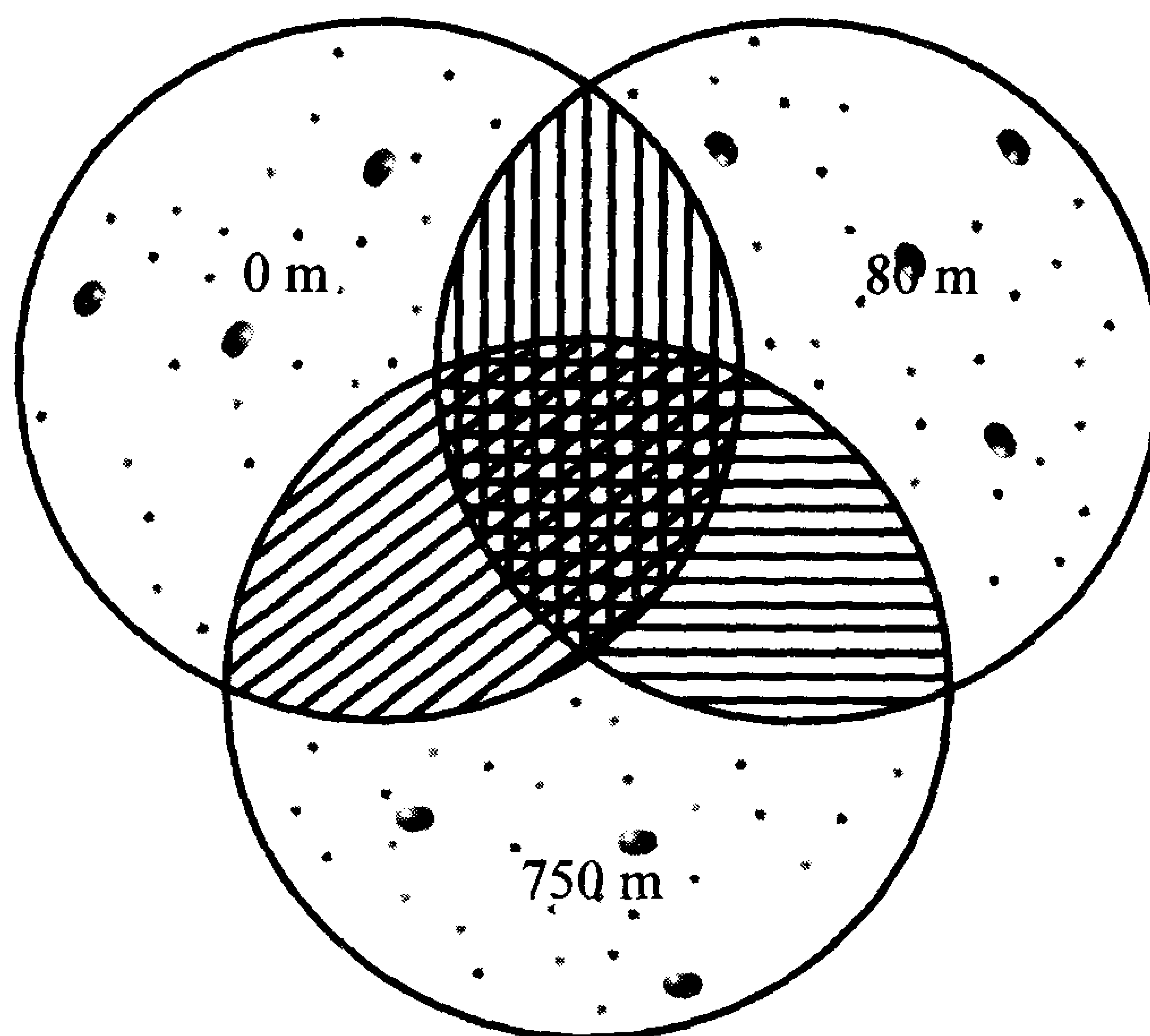


图 8 比较群体基因组学。表中列举了正在研究的几个蒙特利海湾不同水域的 BAC 文库（又见 http://www.tigr.org/tdb/MBMO/）。图中展示了不同水体深度的基因组群体所特有和共有的基因。

比较微生物生态系统基因组学的潜力很大，随着新数据和科技手段的出现，这个新领域必将为我们提供认识世界的全新方法。

结论

对技术杠杆和经济利益方面的认识，继续朝着有利该领域的方向发展，促进微生物基因组学和生态学的研究，包括 DNA 自动测序、生物信息学、蛋白组学、微阵列技术

以及自动化环境感应器技术领域的进步，可能在未来微生物群体基因组学和生态学综合研究中发挥重要作用。在这种数据密集情况下，面临的更大挑战是，按空间和时间分布将这些分散的观察和数据统一起来，在获得大量数据的同时，基因组和生态学理论也逐渐成熟，这将反过来帮助指导未来的研究，从而以便获得更多的信息。

Zucherandl 和 Pauling[3]断言，单个大分子记录着进化史的信息，把这种观点按照逻辑稍稍拓展，就能明白自然基因组也记录着环境的、进化的和生态的历史及其动态变化信息。环境基因组学研究运用免培养方法，这种方法使我们能够研究未能培养微生物大量种类的基因组，为未能培养微生物的基因组分析，提供有关微生物群体内个体生物学特性更为深入的认识。现在，运用微生物群体基因组学，可由基因型对表现型进行预测，扩大对自然环境过程的认识，认识这种真实未经篡改的多样性，可以提高开发利用微生物天然产物和过程的能力，反映基因组群体内部或之间的自然属性，以及它们与环境之间相互关系的基因组模式也会更加明朗。

自然微生物群体基因组学研究刚刚起步，机遇很多，着手研究基因组群体动力学，在不久的将来就会成为现实，基因组群体动力学可能会随着包括基因漂移、重组、分散、竞争、迁移和承袭等机制而波动和发展。经典群体遗传学理论，与微生物群体内部及之间基因组交换和进化的研究相结合形成了一个新领域，在该领域中，微生物生态学和基因组学研究在不久的将来会整合到一起。自然微生物世界为研究和认识基因组进化过程和动力学提供了理想的材料，相反，基因组学为揭示微生物群落结构和动力学基础的基因组进化过程、形态复杂性提供了必要的数据，微生物生态学和基因组学现在必然会协同渗透，并导致产生新理论和加深对周围自然微生物界的认识。

（张利莉，江 昊 译）

参考文献

1. Ward DM. A natural species concept for prokaryotes. Curr Opin Microbiol 1998; 1:271–277.
2. Stackebrandt E, Frederiksen W, Garrity GM, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 2002; 52:1043–1047.
3. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Theor Biol 1965; 8:357–366.
4. Pace NR, Stahl DA, Olsen GJ, Lane DJ. Analyzing natural microbial populations by rRNA sequences. ASM News 1985; 51:4–12.
5. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 1986; 40:337–365.
6. DeLong EF, Wickham GS, Pace NR. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. Science 1989; 243:1360–1363.
7. Amann RI, Ludwig W, Schleifer K-H. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol Rev 1995; 59:143–169.
8. Orphan VJ, House CH, Hinrichs KU, McKeegan KD, DeLong EF. Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. Science 2001; 293:484–487.
9. Orphan VJ, House CH, Hinrichs KU, McKeegan KD, DeLong EF. Multiple archaeal groups

mediate methane oxidation in anoxic cold seep sediments. Proc Natl Acad Sci USA 2002; 99: 7663–7668.

10. Schmidt TM, DeLong EF, Pace NR. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 1991; 173:4371–4378.
11. Saiki RK, Gelfand DH, Stoffel S, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988; 239:487–491.
12. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. Nature 1990; 345:60–63.
13. Pace NR. A molecular view of microbial diversity and the biosphere. Science 1997; 276:734–740.
14. Woese CR. Bacterial evolution. Microbiol Rev 1987; 51:221–271.
15. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 1998; 180:4765–4774.
16. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 2002; 418:630–663.
17. Connon SA, Giovannoni SJ. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol 2002; 68:3878–3885.
18. DeLong EF. Everything in moderation: archaea as "non-extremophiles." Curr Opin Genet Dev 1998; 8:649–654.
19. DeLong E. Archael means and extremes. Science 1998; 280:542–543.
20. DeLong EF, Taylor LT, Marsh TL, Preston CM. Visualization and enumeration of marine planktonic archaea and bacteria by using polyribonucleotide probes and fluorescent *in situ* hybridization. Appl Environ Microbiol 1999; 65:5554–5563.
21. DeLong EF, King LL, Massana R, et al. Dibiphytanyl ether lipids in nonthermophilic crenarchaeotes. Appl Environ Microbiol 1998; 64:1133–1138.
22. Karner MB, DeLong EF, Karl DM. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature 2001; 409:507–510.
23. Lopez-Garcia P, Lopez-Lopez A, Moreira D, Rodriguez-Valera F. Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. FEMS Microbiol Ecol 2001; 36:193–202.
24. Lopez-Garcia P, Moreira D, Lopez-Lopez A, Rodriguez-Valera F. A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. Environ Microbiol 2001; 3:72–78.
25. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 2002; 417:63–67.
26. Moore LR, Rocap G, Chisholm SW. Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. Nature 1998; 393:464–467.
27. Rocap G, Distel DL, Waterbury JB, Chisholm SW. Resolution of Prochlorococcus and Synechococcus ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 2002; 68:1180–1191.
28. DeLong E, Pace NR. Environmental diversity of bacteria and archaea. Systematic Biol. 2001; 50:1–9.
29. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
30. Nelson KE, Paulsen IT, Fraser CM. Microbial genome sequencing: a window into evolution and physiology. ASM News 2001; 67:310–317.
31. Kim U-J, Shizuya H, Dejong P, Birren B, Simon M. Stable propagation of cosmid sized human DNA inserts in an F-factor based vector. Nucleic Acids Res 1992; 20:1083–1185.
32. Shizuya H, Birren B, Kim UJ, et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci

USA 1992; 89:8794–8797.
33. Doolittle RF. Microbial genomes opened up. Nature 2002; 392:339–342.
34. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001; 409:1007–1011.
35. Andersson SG, Zomorodipour A, Andersson JO, et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 1998; 396:133–140.
36. Gardner MJ, Tettelin H, Carucci DJ, et al. The malaria genome sequencing project: complete sequence of *Plasmodium falciparum* chromosome 2. Parassitologia 1999; 41:69–75.
37. Tamas I, Klasson L, Canback B, et al. Fifty million years of genomic stasis in endosymbiotic bacteria. Science 2002; 296:2376–2379.
38. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. Nature 2000; 407:81–86.
39. Béjà O, Aravind L, Koonin EV, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 2000; 289:1902–1906.
40. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bacteriol 1996; 178:591–599.
41. Shizuya H, Kouros-Mehr H. The development and applications of the bacterial artificial chromosome cloning system. Keio J Med 2001; 50:26–30.
42. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 1998; 180:5003–5009.
43. Béjà O, Koonin EV, Aravind L, et al. Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. Appl Environ Microbiol 2002; 68:335–345.
44. Cai W, Jing J, Irvin B, et al. High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. Proc Natl Acad Sci USA 1998; 95:3390–3395.
45. Lanoil BD, Carlson CA, Giovannoni SJ. Bacterial chromosomal painting for *in situ* monitoring of cultured marine bacteria. Environ Microbiol 2000; 2:654–665.
46. Suzuki M, Rappe MS, Giovannoni SJ. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. Appl Environ Microbiol 1998; 64:4522–4529.
47. Marsh TL, Saxman P, Cole J, Tiedje J. Terminal restriction fragment length polymorphism analysis program, a Web-based research tool for microbial community analysis. Appl Environ Microbiol 2000; 66:3616–3620.
48. Béjà O, Suzuki MT, Heidelberg JF, et al. Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature 2002; 415:630–633.
49. Song J, Dong F, Lilly JW, Stupar RM, Jiang J. Instability of bacterial artificial chromosome (BAC) clones containing tandemly repeated DNA sequences. Genome 2001; 44:463–469.
50. Rondon MR, Raffel SJ, Goodman RM, Handelsman J. Toward functional genomics in bacteria: analysis of gene expression in *Escherichia coli* from a bacterial artificial chromosome library of *Bacillus cereus*. Proc Natl Acad Sci USA 1999; 96:6451–6455.
51. Rondon MR, August PR, Bettermann AD, et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 2000; 66:2541–2547.
52. Béjà O, Suzuki MT, Koonin EV, et al. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ Microbiol 2000; 2:516–529.
53. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 2001; 67:4399–4406.

54. Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. Proteorhodopsin phototrophy in the ocean. Nature 2001; 411:786–789.
55. Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ. Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. Appl Environ Microbiol 1998; 64:3075–3078.
56. Schleper C, Swanson RV, Mathur EJ, DeLong EF. Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 1997; 179: 7803–7811.
57. Preston CM, Wu KY, Molinski TF, DeLong EF. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen nov, sp nov. Proc Natl Acad Sci USA 1996; 93: 6241–6246.
58. Short JM. Recombinant approaches for accessing biodiversity. Nat Biotechnol 1997; 15:1322–1323.
59. Majernik A, Gottschalk G, Daniel R. Screening of environmental DNA libraries for the presence of genes conferring Na(+)(Li(+))/H(+) antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. J Bacteriol 2001; 183:6645–6653.
60. Henne A, Daniel R, Schmitz RA, Gottschalk G. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl Environ Microbiol 1999; 65:3901–3907.
61. Henne A, Schmitz RA, Bomeke M, Gottschalk G, Daniel R. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. Appl Environ Microbiol 2000; 66:3113–3116.
62. Cottrell MT, Moore JA, Kirchman DL. Chitinases from uncultured marine microorganisms. Appl Environ Microbiol 1999; 65:2553–2557.
63. Richardson TH, Tan X, Frey G, et al. A novel, high performance enzyme for starch liquefaction: discovery and optimization of a low pH thermostable alpha-amylase. J Biol Chem, 2002.
64. MacNeil IA, Tiong CL, Minor C, et al. Expression and isolation of antimicrobial small molecules from soil DNA libraries. J Mol Microbiol Biotechnol 2001; 3:301–308.
65. Brady SF, Chao CJ, Handelsman J, Clardy J. Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. Org Lett 2001; 3:1981–1984.
66. Hess WG, Rocap G, Ting CS, et al. The photosynthetic apparatus of *Prochlorococcus*: insights through comparative genomics. Photosynthesis Res 2001; 70:53–71.
67. Ting CS, Rocap G, King J, Chisholm SW. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol 2002; 10: 134–142.
68. Spratt BG, Hanage WP, Feil EJ. The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr Opin Microbiol 2001; 4:602–606.
69. Smith JM, Feil EJ, Smith NH. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 2000; 22:1115–1122.
70. Feil EJ, Spratt BG. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 2001; 55:561–590.
71. Curtis TP, Sloan WT, Scannell JW. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA 2002; 99:10,494–10,499.
72. Kolber ZS, Plumley FG, Lang AS, et al. Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. Science 2001; 292:2492–2495.
73. Kolber ZS, Van Dover CL, Niederman RA, Falkowski PG. Bacterial photosynthesis in surface waters of the open ocean. Nature 2000; 407:177–179.
74. Field KG, Gordon D, Wright T, et al. Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. Appl Environ Microbiol 1997; 63:63–70.

24

基因组学在生物催化和生物降解中的应用

Lawrence P. Wackett

“人类两种对立的冲动驱使我们赖以生存的这个世界，在两个极端中摇摆——要么是美丽的花园，要么是荒芜的沙漠。我们的将来与人类的创造力紧紧地联系在一起。”

Mihaly Csikszentmihalyi

引言

在地球上所有主要生物群体中，原核生物具有最多样化的代谢方式[1]，这与它们种系发生（phylogenetic）的多样性，以及在地球上很多独特小生境中的生存能力有关。原核生物是生物圈中碳元素的主要回收者，它们惊人的个体数目（5×10^{30}）以及超过世界上所有绿色植物的生物量，就是它们卓有成效的具体见证[2]。生物催化（biocatalysis）和生物降解（biodegradation），具体体现了微生物代谢的天然多样性。

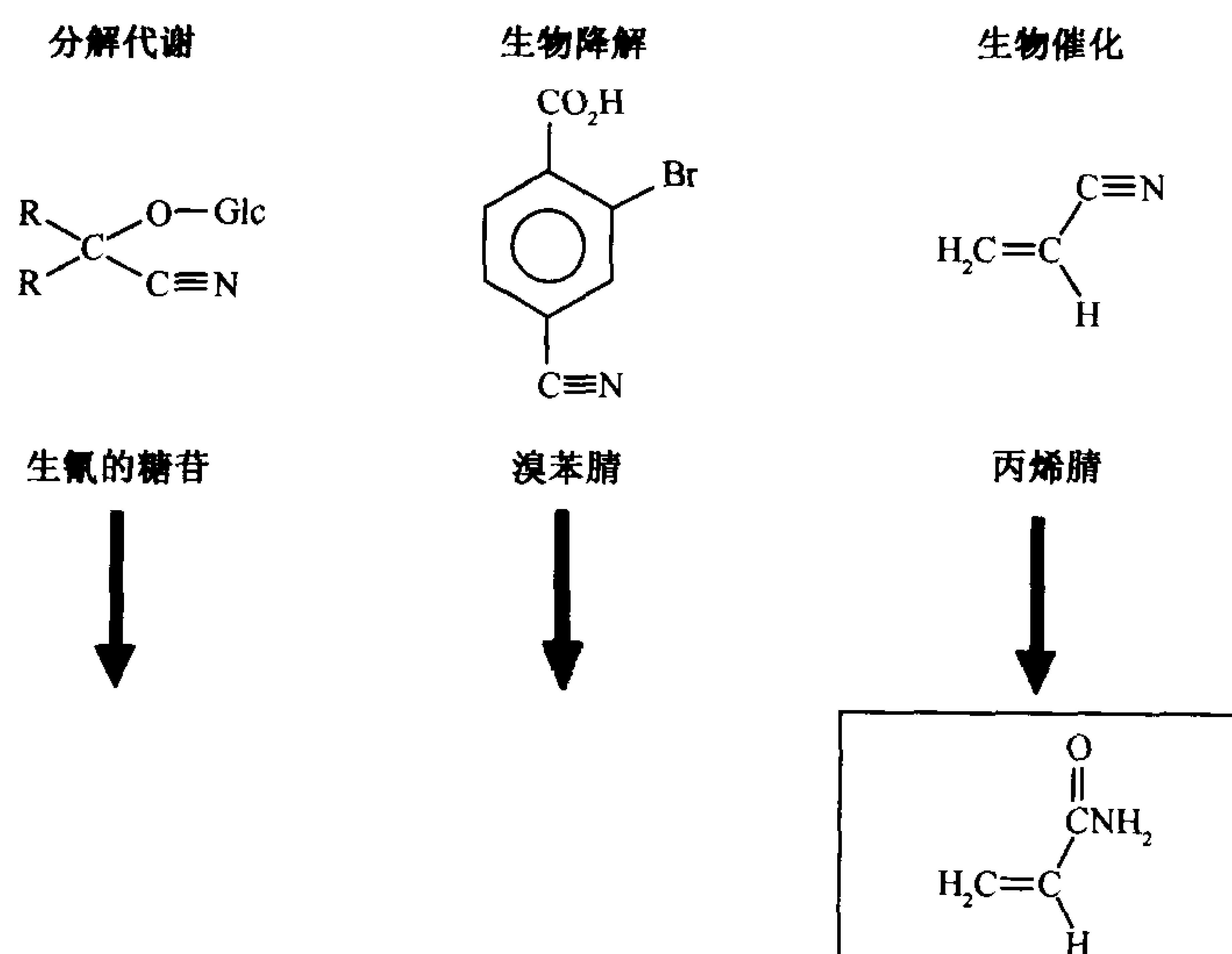


图 1　微生物的酶对不同化合物中的氰（cyano）或腈（nitrile）功能基团的水合作用，可根据该反应对微生物或人类有意义的分解代谢、生物降解或生物催化。丙烯酰胺主要是由生物催化生产的商业化学产品。

生物降解是指微生物对有机物的降解。人类观察到这一现象已有几千年了，如由腐

生真菌而引起林木的自然腐烂，直到近期，才对林木腐烂的生物化学反应——代谢反应(metabolism)渐渐有所认识[3,4]。降解化合物并从中获取能量的代谢反应叫异化作用或分解代谢（catabolism)。随着对微生物异化途径了解的逐渐深入，有时会利用这些途径来清理有毒有害的废料，这一过程称为生物治理（bioremediation)。

生物催化与生物降解相关，在通常情况下它是指利用微生物的异化反应和生物合成来生产商业化学物质，例如，抗生素生产就是发酵、酶工程和有机合成等手段的综合利用。工业上对生物催化的利用越来越多，部分情况是由于发现了新的生物降解途径，试举一例，腈水合酶（nitrile hydratase）用于把丙烯腈（acrylonitrile）转化为丙烯酰胺(acrylamide) （图 1)。由于酶反应干净并且转化效率高，日本日东化学公司（Nitto Chemical Company）在此基础上建立了大规模生物工程反应体系[5]，细菌合成腈水合酶用来生物降解腈化物以及由此产生的氰化物，从而去除这些物质的毒性；植物生产这些物质来抵抗害虫（图 1)[6]；有些细菌的腈水合酶还可降解人工合成的除草剂，如溴苯腈（bromoxynil)(图 1)。因此，自然进化的酶可作用于天然产物、生物降解污染物以及生物合成工业的化学产品。

生物降解

原核生物的生物降解

生活在自然界土壤和水体中的细菌，一般都具备足够广泛的分解代谢能力，以使它们能在激烈竞争的环境中摄取生存所必需的碳、氮、磷以及其他营养，因此，土壤微生物就比肠道微生物和病原微生物具有更广泛的生物降解途径。大肠杆菌（*Escherichia coli*）在土壤和水体中难以生存，却能在动物肠道中快速生长，肠道提供相当单一的限制性的营养。有观察显示，大肠杆菌能分解一些芳香类有机酸，这并不奇怪，因为这些化合物是由芳香类氨基酸和植物天然产物降解而来，而这些物质存在于大肠杆菌所生活的肠道环境中[8,9]。

在研究得比较清楚的原核生物中，是否有些微生物不主要依赖那些已知的生物降解途径？可联想到古生菌（Archaea)，一个主要的生命形式分支，这群微生物在过去 10 年中已研究得很深入，几个全基因组测序项目已完成，包括能在高盐或高温条件下旺盛生长的极端微生物（extremophiles)。在古生菌中的一个主要代谢类群称为产甲烷菌(methanogens)，它们把二氧化碳和乙酸转化为甲烷，从中获取能量并释放甲烷到大气层中。

曾用产甲烷菌纯培养物研究还原氯代脂肪族化合物（chlorinated aliphatic compounds）的能力，一般都认为菌体通常不进行这些反应[10]，而早就发现氯代甲烷对产甲烷菌有毒，它们与正常生理物质争夺钴胺素（cobalamin）（维生素 B_{12}）的活性中心，或在它们还原后立即生成活性炭烯（reactive carbene）中间产物。总之，在分解各种有机物时，产甲烷菌不起显著作用，与此类似，嗜盐古生菌也没有广泛异化能力。尽管曾报道它们中有的能降解脂肪族碳水化合物，但一般认为它们能利用碳源的种类比较有限[11]，此外，盐浓度增加可导致碳水化合物的生物降解减弱，尤其是在降解石油泄漏后的芳香族碳水化合物时，菌群也由原来的细菌转变为古生菌。

尽管如此，异化能力广泛存在分类树中的生命体中（表 1），例如，表 1 中列举一些属（genera）的细菌，都具有显著的异化代谢能力，已被用在工业化生物降解和生物催化中，这些细菌大多属低 G+C 含量或高 G+C 含量的革兰氏阳性菌，或多样菌（Proteobacteria）。

表 1　明尼苏达大学生物催化/生物降解数据库（UM-BBD）根据细菌分解代谢能力收集一些细菌属的资料[a]

高 G+C 含量革兰氏阳性	低 G+C 含量革兰氏阴性	多形杆菌门 α	β	γ	δ/ε	噬纤维菌目绿色硫细菌	绿色非硫细菌
节杆菌属	芽胞杆菌属	农杆菌属	无色杆菌属	不动杆菌属	脱硫弧菌属	**黄杆菌属**	脱卤拟球菌
短杆菌属	**梭菌属**	屈曲杆菌属	产碱菌属	气单胞菌属			
棒状杆菌属	脱硫杆菌属	短波单胞菌属	固氮弧菌属	固氮菌属			
棒杆菌属	真杆菌属	*Chelatobacter*	**伯克霍尔德菌属**	肠杆菌属			
脱卤杆菌	葡萄球菌属	*Hypomicrobium*	丛毛单胞菌属	**埃希氏菌属**			
奴卡菌属		甲基杆菌属	*Hydrogenophyga*	克雷伯氏菌属			
红球菌属		副球菌属	罗尔斯顿氏菌属	甲基杆菌属			
链霉菌属		红杆菌属	*Thavera*	甲基球菌属			
地杆菌属		鞘氨醇单胞菌属	硫杆菌属	莫拉氏菌属			
				假单胞菌属			

[a]属名黑体表示该属的资料相当丰富。

虽然很多细菌能在生物降解中起作用，有的菌可能更重要。早在 1926 年，den Dooren de Jong[12]报道一种假单胞菌（*Pseudomonas*）可以降解上百种有机化合物，包括烷类和芳香族环类化合物，至今仍有人认为细菌对这些底物的偏爱是“奇特的（exotic)”。但是，烷类和芳香族环类化合物普遍存在，它们在地球上以相当大的数量存在了数百万年了[13]，微生物难以抵挡这些含丰富热量化合物的诱惑。

生物降解起重要作用的一些细菌基因组学

基因组学才刚开始影响生物降解研究的主流（表 2），头几年的基因组测序几乎全都围绕着病原细菌，因为第一个基因组测序项目是由美国国立卫生研究院（U.S. National Institute of Health）资助，其目的就是要直接服务于人类健康。但是，现在对种系多样原核生物测序的资助已大大增加，在公共范围内，约有 500 个细菌基因组测序已经完成或正在进行，预计几百个全基因组序列很快就会面世，其中包括相当一部分是种系多样的土壤微生物（表 2）。

基因组学将不断加深对土壤中细菌代谢活动的认识，例如，有关枯草芽孢杆菌基因组测序的文章说，该菌有能降解某些植物天然产物的基因[14]，对此令人惊讶，因为，此前这些基因是在与该菌分类地位完全不同的一株革兰氏阴性土壤细菌中发现的。然而，以核糖体核糖核酸（ribosomal ribonucleic acid，rRNA）作指示物的研究显示，每克土壤中含有多达一万多种不同细菌[15]，而黄酮类(flavonoid)物质可占某些植物叶片

表 2 公共领域中与生物催化和生物降解有关原核生物基因组计划[a]

物种	基因组大小/kb	在生物催化和生物降解中的应用
已完成基因组		
不动杆菌（*Acinebacter* sp）ADP1 ATCC 33305	3583	烷烃/安息香酸盐代谢模型
丙酮丁醇梭菌（*Clostridium acetobutylicum*）ATCC 824 D	4100	丙酮/丁醇发酵
谷氨酸棒杆菌（*Corynebacterium glutamicum*）	3309	谷氨酸发酵
谷氨酸棒杆菌（*Corynebacterium glutamicum*）ATCC 13032	3309	氨基酸发酵
耐辐射异常球菌（*Deinococcus radiodurans*）R1	3284	放射性环境中的生物降解
阿维链霉菌（Streptomyces avermitilis）MA-4680	8700	抗生素生产
天蓝色链霉菌（Streptomyces coelicolor）A3（2）	8667	抗生素生产
多态链霉菌（*Streptomyces diversa*）	不详	抗生素生产
运动发酵单胞菌（*Zymomonas mobilis*）ZM4	2052	乙醇/山梨糖醇发酵
运动发酵单胞菌（*Zymomonas mobilis*）	1833	乙醇/山梨糖醇发酵
正在测定基因组		
洋葱伯克霍尔德菌（假单胞菌）(*Burkholderia*（*Pseudomonas*）*cepacia*）J2315	7600	杀虫剂的生物降解
有效棒杆菌（*Corynebacterium efficiens*）YS-314T	3140	谷氨酸发酵
热产氨棒杆菌（*Corynebacterium thermoaminogenes*）FERM9246	不详	氨基酸发酵
乙烯脱卤拟球菌（*Dehalococcoides ethenogenes*）	1500	溶剂的还原性脱卤
Desulfitobacterium hafniense	4600	还原性脱卤
金属还原地杆菌（*Geobacter metallireducens*）	6800	有毒金属的还原/固定
硫还原地杆菌（*Geobacter sulfurreducens*）	2500	金属还原
欧洲亚硝化单胞菌（*Nitrosomonas europaea*）ATCC 25978	2980	氮循环；溶剂氧化
荧光假单胞菌（*Pseudomonas fluorescens*）Pf0-01	3500	多种生物降解能力
荧光假单胞菌（*Pseudomonas fluorescens*）SBW25	6600	多种生物降解能力
恶臭假单胞菌（*Pseudomonas putida*）KT2440	6100	多种生物降解能力
恶臭假单胞菌（*Pseudomonas putida*）PRS1	6100	多种生物降解能力
耐金属罗尔斯顿菌（富营养）(*Ralstonia metallidurans*（eutropha））CH34	3000	重金属抗性
红球菌（*Rhodococcus* sp）I24	5487	茚的生物转化
红球菌（*Rhodococcus* sp）RHA1	不详	多氯联苯的生物降解
沼泽红假单胞菌（*Rhodopseudomonas palustris*）CGA009	5460	光养型降解芳香族化合物
嗜芳香物鞘氨醇单胞菌（*Sphingomonas aromaticivorans*）F199	3800	多种芳香族化合物的降解
Streptomyces ambofaciens	8000	抗生素生产

[a]数据取自《基因组在线数据库》(Genomes Online Database, GOLD)，2002 年 6 月 12 日。

生物量的27%，这些叶片物质可为土壤提供主要的新增碳源[6]。因此，在该类植物生长的温带土壤中，这些有机物的广泛存在，使分类地位迥然不同的微生物含有分解这些物质的基因，而这些基因不大可能在温泉、极地和沙漠中发现。因此，基因簇不仅把单个微生物相互联系在一起，而且也可以把它们及其周围具有复杂生物特性的生存环境联系在一起，这就开始形成全球基因组组成（global genomic composition）研究的一个分支方向。尽管这种研究永远难以完成，因为这是把相关信息放在生态研究领域中加以考虑，有助于给基因组学下一个全面广泛的定义。

起异化作用质粒的基因组学

许多土壤微生物都有染色体外脱氧核糖核酸（deoxyribonucleic acid，DNA）称为质粒（plasmid）的因子。随着基因组测序的快速扩展，在细菌中越来越频繁地发现大量的小DNA因子，这使得质粒组成成分的概念变得模糊了。历史上，质粒于1952年在大肠杆菌中首先报道[16]，从那以后，相继发现质粒可以携带抗生素抗性基因、致病基因以及代谢各种化学物质的基因。很多质粒还携带能在细菌接合过程中促进自身转移的基因，寄主广泛的质粒能在不同属的细菌间转移和复制。在这个意义上，质粒对异化作用基因在环境中的转移起非常重要的作用，有的质粒可以编码利用辛烷（octane）、甲苯（toluene）、樟脑（camphor）、萘（naphthalene）、烟碱（nicotine）、对甲苯磺酸（*p*-toluenesulfonic acid）以及2,4-二氯-乙酰苯酯（2,4-dicholorophenoxyacetate）的基因[17~21]。

在假单胞菌等土壤细菌中，质粒DNA可占总DNA的相当一部分，例如，分离的可以降解除草剂莠去津（atrazine）的一株假单胞菌，含有多达5个、总长约1Mb、具有不同异化作用的质粒[22]。

基因组学研究方法毫无疑问地增加对质粒结构和进化关系的了解，但是到目前为止，只测序了几百个质粒，其中大多数是分子生物学所用的小载体质粒，或是病原细菌中的抗生素抗性质粒。表3列举了已完成或正在测序有分解代谢作用的一些质粒，对嗜芳香物鞘氨醇单孢菌（*Sphingomonas aromaticivorans*）菌株F199的分解代谢质粒pNL1已完全测序，它含有能代谢联苯（biphenyl）、萘、间二甲苯（*m*-xylene）和对甲酚（*p*-cresol）的酶[23]，该菌株的全基因组测序也已完成。

表3　代谢质粒DNA的全序列

物种	质粒	大小/kb	功能
已完成的			
嗜芳香物鞘氨醇单孢菌（*Sphingomonas aromaticivorans*）F199	pNL-1	186	芳香族化合物的降解
假单胞菌（*Pseudomonas* sp）ADP	pADP-1	107	莠去津的代谢
恶臭假单胞菌（*Pseudomonas putida*）TOL	pWWO	117	甲苯/二甲苯的代谢
正在测序的			
假单胞菌（*Pseudomonas* sp）ND6	pND6-1	102	萘的代谢
嗜烟碱节杆菌（*Arthrobacter nicotinovorans*）	pAO1	160	烟碱（尼古丁）的代谢

对假单胞菌菌株 ADP 的代谢质粒 pADP-1 的测序最近已经完成[24]（图 2），注释显示与该质粒复制、转移和维持（maintenance）有关的基因，几乎与质粒 pR751 的类似基因完全一致，pR751 来自产气肠杆菌（*Enterobacter aerogenes*）的 IncPβ 质粒[25]。降解莠去津整个代谢途径所需的各种 *atz* 基因都定位于 pADP-1 质粒上，但它们并不相邻排列形成一个类似操纵子的结构（图 2），事实上，编码莠去津降解途径中前三个反应酶的基因都各自分散在质粒上，它们两端都紧邻插入序列元件，这三个基因似乎是持续表达的。其中，*atzC* 的 G+C 百分比为 44%，比 *atzA*（58%）和 *atzB*（61%）的低，这表明，这几个 *atz* 基因是不久前被一个有广泛寄主的质粒骨架俘获而产生的一个新质粒，这个新质粒使假单胞菌菌株 ADP 能以莠去津为唯一氮源而生长。

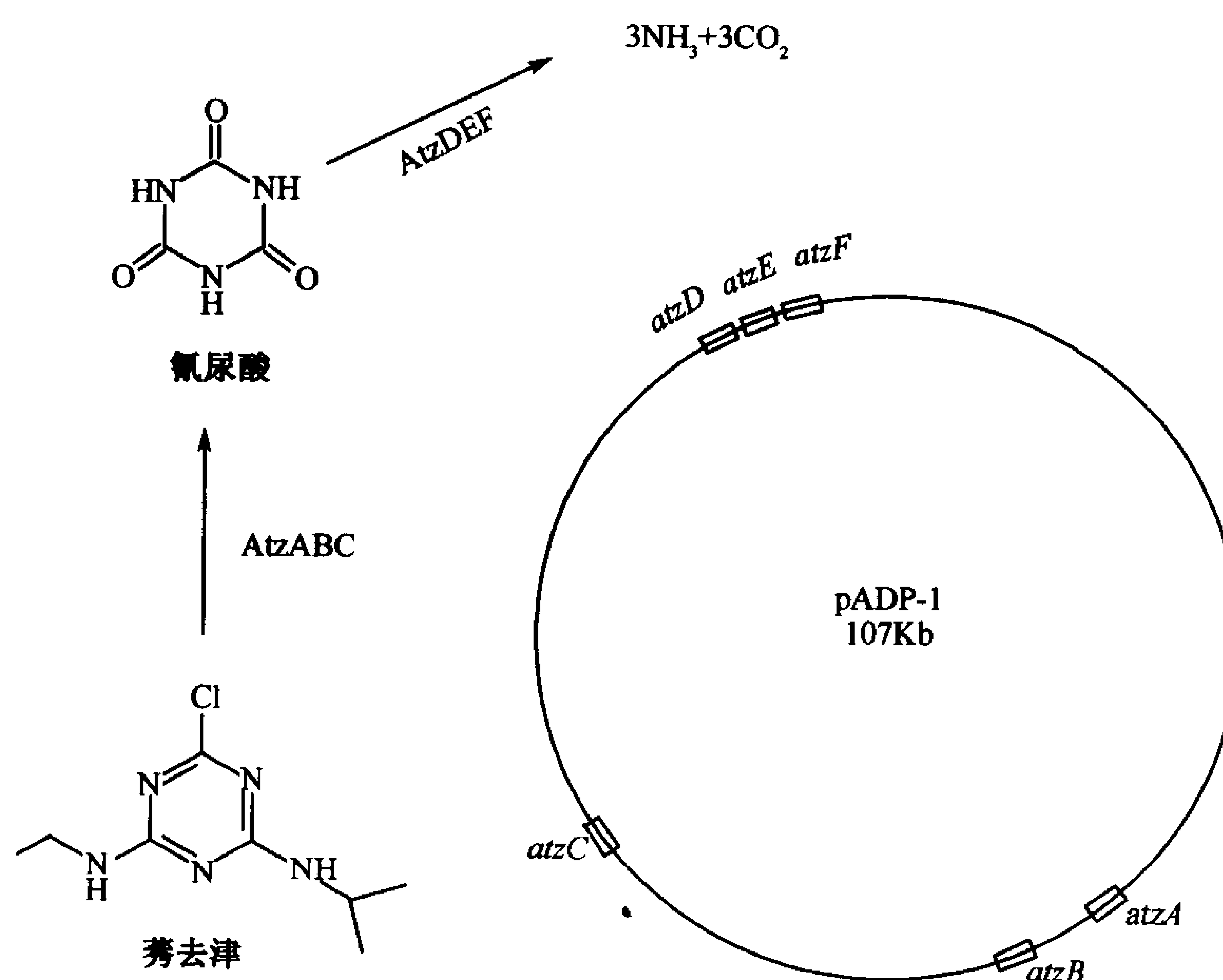


图 2 假单胞菌菌株 ADP 和质粒 pADP-1 对莠去津分解代谢途径。质粒 pADP-1 含有代谢莠去津酶的编码基因。

另一方面，编码莠去津代谢途径中另外三个酶基因却聚在一起，并且是协同控制的[24]。据推测，莠去津代谢途径可以分成“上（upper）”、“下（lower）”两部分，“上”途径的酶是近期才演化成能降解人工合成除草剂均三嗪（*s*-triazine）的活性。最近一项研究与此观点一致：新发现的脱氨酶 TriA 与莠去津氯水解酶（chlorohydrolase）AtzA 同源，它们的序列有 98% 一致性，但是 TriA 是个脱氨酶，它对莠去津几乎没有任何脱氯酶活性，这表明，几个氨基酸的改变可以使 TriA 酶转化为 AtzA 酶。在实验室通过 DNA 改组（DNA shuffling），再筛选改变了催化活性的重组蛋白，也可以达到同样的效果[26]。

用作生物治理的“工程”菌——耐辐射异常球菌的基因组学

为了使生物治理应用得更广泛，期望给微生物设计一些生物降解能力，以使其具有某些独特性能，适用于某些特定场合。例如，改变耐辐射异常球菌（图 3）的代谢途

径，使它能转化含放射性同位素的有机废料，由于放射性同位素集中用在民用核反应堆和军用核弹头上，这类废料保存在美国能源部（US Department of Energy，DOE）所属的许多地方以及其他国家的相应地点。

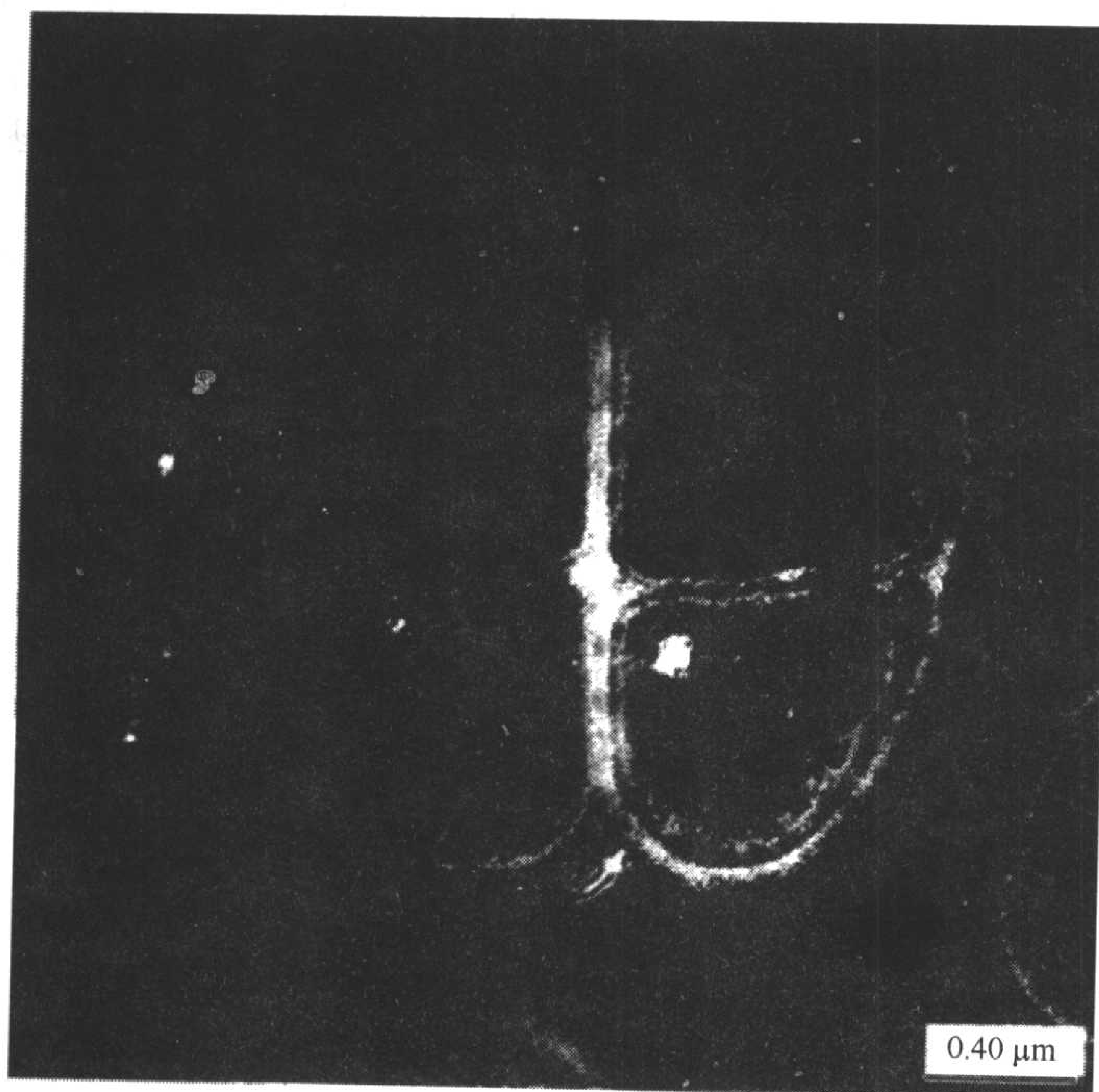


图 3　耐辐射异常球菌菌株 R1 超薄切片的电子显微照片，其细胞呈现典型的四联体排列。

20 世纪 90 年代初，随着冷战的结束，美国能源部启动了一个大项目，其目的是安全保存和净化 7000 个地点的三百万立方米的混合废料[27]，在某些地点，放射性废料正在向周围土壤和地下水泄漏[28]，据估计，被污染土壤有四千万立方米，被污染地下水有四万亿（4 trillion）升，将它们在 10 年内净化的费用预计高达 600 亿美金，为了降低治理费用，美国能源部支持就地生物治理（*in situ* bioremediation）的措施。

美国能源部管辖的许多污染点的电离辐射强度，对表 2 中所列微生物都是致死剂量，因此，有必要用抗辐射的细菌来进行生物治理。耐辐射异常球菌就能忍耐这种辐射的程度，这种菌原来是从受过辐射的罐头肉中分离的，它能耐受超过 15 000Gy（Gray，戈端）的强烈电离辐射[29]，对其他损伤 DNA 的条件也具有很强的抗性，如干旱、紫外线照射以及氧化剂等。它的这些特点是一种综合表现型（a complex phenotype），是由从酶活到基因组结构等一系列因素而决定的，这就使将敏感菌构建为抗辐射菌的策略不那么诱人，相反，注意力却集中在将耐辐射异常球菌构建为有生物降解能力的菌，这就促进了耐辐射异常球菌基因组的测序和注释。基因组数据很快弄清了该菌本身具有的代谢途径，并提供了一张代谢路线图，可帮助设计给现有途径供应中间产物的一些补充代谢途径（complementary metabolic pathway）。

耐辐射异常球菌菌株 R1 的基因组由四个分别长为 2649 kb、412 kb、177 kb 和

45 kb的复制元（replicon）组成[30]，显然，两个最大复制元含有必需基因，而所有四个复制单元都稳定存在，基因组注释显示，基因组 91％是编码区，共编码 3187 个可读框(open reading frame)，有 69％可读框与数据库中的序列相吻合，对这些数据的进一步分析才弄清了这一微生物的代谢途径。

耐辐射异常球菌能编码全套糖酵解（glycolysis）、糖异生（gluconeogenesis）、磷酸戊糖旁路（pentose phosphate shunt）、三羧酸循环（tricarboxylic acid cycle）、乙醛酸旁路(glycoxylate shunt）等代谢途径的基因。乙醛酸旁路在许多原核生物中不存在，但它却在耐辐射异常球菌中能强烈表达[31]，耐辐射异常球菌极其缺乏代谢有机污染物（如芳香族碳水化合物）的酶，因此，天然耐辐射异常球菌，不可能用来降解美国能源部的有机污染物，在室内实验也证明，野生型耐辐射异常球菌 R1 缺乏生物降解表现型（S. McFarlan，未发表数据），这为构建能降解有机和无机毒性化合物的耐辐射异常球菌的代谢工程菌提供了有价值的材料。

耐辐射异常球菌的优点是，很容易转化进外源 DNA，而载体也能在该菌体内复制[29]，此外，Daly 和同事们还设计了一种 DNA 盒式系统（DNA cassette system)[29]，他们把抗生素抗性标记和一系列重复序列放在外源 DNA 的左右两侧，这些重复序列可以和耐辐射异常球菌的基因组重组，在菌的生长期中，不断增加培养基中抗生素浓度，DNA 框就可以在基因组中不断扩增，从而提高基因的拷贝数。用这种方法，已经得到高达 200 个拷贝数的 DNA 框，假设一个拷贝 *mer* 操纵子（mercury resistance operon，抗汞操纵子）约 20 kb，那么，200 个拷贝相当于把耐辐射异常球菌的整个基因组扩大了一倍。

通过这些途径，已经把 *mer* 操纵子和甲苯代谢基因克隆到耐辐射异常球菌中，并在菌体内扩增和表达[32,33]，*mer* 操纵子编码一种可溶的汞还原酶和几种汞离子转运蛋白[34]。虽然 Hg^{2+} 对细菌有高毒性，但是，一些细菌可以把 Hg^{2+} 转运到细胞内还原成金属汞（Hg^0），金属汞的毒性小得多，而且，有足够挥发性使它很容易从细胞中逃逸出去。野生耐辐射异常球菌对汞没有抗性，基因组注释也未发现它有能解除汞毒性的基因，用上述方法，把大肠杆菌的整个汞操纵子和一个氨苄抗性基因一起克隆到耐辐射异常球菌中。

正如所料，提高培养基中氨苄青霉素浓度，每个细胞 *mer* 基因的拷贝数也相应增加，这就提高了菌体对汞离子的抗性。而且，含有 *mer* 操纵子和与甲苯氧化有关酶的重组耐辐射异常球菌，但可以在对野生菌有毒的高汞离子浓度下氧化甲苯。美国能源部管辖的污染点均含有相当量的汞，因此，对准备在这些点实施生物治理的细菌，*mer* 基因是必不可少的。野生耐辐射异常球菌可以还原某些金属，如铀（uranium）和锝(technitium)，它们是能源部管辖污染点重要的放射性核素（radionuclide)[35]，还原这些金属离子降低它们在土壤中的迁移能力，从而减少它们污染周围环境的概率。

用生物方法还原金属要求有可氧化底物存在，如果这些底物本身是废料堆放点的污染物那就再好不过了，许多污染点存在燃料类碳水化合物，其中甲苯占绝大多数，在这种情况下，编码甲苯加双氧酶（toluene dioxygenase）的 *tod ABC1C2* 基因，克隆到耐辐射异常球菌中[32]，即使在电离辐射的情况下，重组菌也能氧化甲苯、氯苯（chlorobenzene）以及三氯乙烯（trichloroethylene)。然而，最初氧化碳水化合物的反应却不能为

生物还原反应提供电子，为此，*todABC1C2DE* 和 *xylFJQK* 基因簇又接着克隆到耐辐射异常球菌中（H. Brim 等，未发表数据），在这些基因的共同作用下，甲苯分解为丙酮酸和乙酸，乙酸是耐辐射异常球菌生长的良好碳源，这与基因组注释结果一致，并显示乙醛酸旁路在这一微生物中是一条高表达的代谢途径。

生物催化

原核生物在生物催化中的作用

1917 年开始用丙酮丁醇梭菌（*Clostridium acetobutylicum*）发酵玉米淀粉生产丙酮，这是用微生物进行生物催化的一次历史性伟大胜利，其工艺流程的开发在很大程度上要归功于 Chaim Weizmann，他为英国在第一次世界大战中打败德国做出了巨大贡献[36]。当时，英国需要丙酮制炸药，战争开始后，来自德国化工厂的丙酮供应被切断，当时的丙酮发酵规模庞大，地域也不仅局限在英格兰，1918 年加拿大有 22 个容积达 3 万加仑的发酵罐在运转[37]。但是，战后世界市场重新恢复，由石油生产丙酮的工艺又占了主流。现在丙酮-丁醇梭菌菌株 ATCC 824D 的基因组测序已经完成（表 2），很快会对丙酮-丁醇发酵的代谢机制有更深入了解，这就可能使生物催化原理用于生产有机溶剂的方法更具有经济上的竞争力。

传统梭菌发酵丙酮的方法在几个方面上预示了现今生物催化的发展，用生物工程生产化学物质的根本是用生物量作投料（feedstock），而不是用石油，用生物或化学的方法把各种来源的生物量转化成葡萄糖，再用它作底物供微生物生产化学物质，有时用酶代替微生物。当前比较重要的生物转化例子是用 α 淀粉酶（α-amylase）把玉米淀粉转化为葡萄糖，再用葡萄糖异构酶转化为果糖高含量的糖浆，一种广泛应用的食品增甜剂(sweetener)[1]。

化工界正在预期由石油投料到生物量投料的重大转变，纤维素是地球上主要生物高聚体（biopolymer），它可以成为重要的生产原料。与淀粉类似，纤维素也可在纤维素酶的作用下酶解为葡萄糖，许多原核生物和真菌都能分泌胞外纤维素酶来提供其生长所需葡萄糖。例如，绿色木霉（*Trichoderma viride*）是已知能最有效分泌胞外纤维素酶的微生物之一[38]。在工业上，从瑞氏木霉（*Trichoderma reesei*）中提取纤维素酶用来发酵乙醇。热纤梭菌（*Clostridium thermocellum*）这样的嗜热细菌的纤维素酶也是研究的对象，热纤梭菌的纤维素酶包装在一个名为纤维体（cellulosome）的、由至少 15 个蛋白质组成的复合体中[39]。

水解生物高聚体产生葡萄糖和其他糖类作为发酵工艺的原料，来生产一些重要的工业化合物，大多数细菌都能利用葡萄糖。因此，可以开发不同微生物的独特生化性能，把它们用在以葡萄糖为原料的发酵过程中来生产期望的化学终产物。利用不同的专业化微生物生产不同的化学产品，基因组学将成为最有效利用这些微生物的重要工具。在抗生素生产上有重要价值的天蓝色链霉菌（*Streptomyces coelicolor*）已被测序，另外被测序的还有棒杆菌（*Corynebacteria*）属的几个菌株，它们中有的正用于氨基酸发酵（表 2)。像这样的其他很多菌株在工业界已经测序，这种趋势无疑还在增长，基因组学将有助于缩短从发现某一菌株到把它用于发酵生产化学产品所需的时间。

生物催化有重要意义细菌的基因组学

在用于公共领域的原核生物基因组测序的背景下，有一大块丑陋的私有原核生物基因组学，这里很大一部分是与工业生物转化有关的菌株。企业利用基因组学研究一些重要菌株的整体代谢活动，这些菌株可用来生产抗生素、氨基酸、生物杀虫剂、维生素、有机酸以及酒精。如果微生物已经用来生产产品，可寄希望基因组学来提高单位体积的产量，从而优化已很赚钱的工艺。幸运的是，虽然我们无法得到那些私有信息，但是公共所有的、与那些重要生物催化相关一些菌株的基因组数据却越来越多，表 2 列举了这样一些菌株。

基因组学研究可以影响生物催化领域的多个方面，例如，棒杆菌属某些菌株生产氨基酸（比如赖氨酸）最重要，大量赖氨酸用于动物饲料中，因为赖氨酸经常是饲养动物饮食中的限制因子，添加赖氨酸可明显增加动物体重。目前，谷氨酸棒杆菌（*Corynebacterium glutamicum*）测序正在进行，它的数据将会公开（表 2)。

红球菌（*Rhodococcus*）在一些生物转化和生物降解过程中起重要作用，它们可以降解多种底物：小分子气体化合物、燃料添加剂甲基叔丁醚（methyl *tert*-butyl ether）、萜类化合物（terpenes），还有除草剂莠去津。此外，对红球菌的兴趣是在生物工程中的应用，实际上，用红球菌的一个菌株把丙烯腈转化为丙烯酰胺是生物催化领域产量最大的工艺过程之一[5]（图 1)。工业界生产 4 亿磅丙烯酰胺，用铜做催化剂的传统化学合成法，却因催化剂带来的高成本和产品杂质所困扰。用红球菌产生的腈水合酶生产就可避免这些问题，并适合于丙烯酰胺的大量生产，日本日东化学公司完全利用微生物催化剂开发了一套商业生产线。

红球菌还能用作生物催化剂对化石燃料进行脱硫处理[40]，化石燃料含有不同程度的有机硫，燃烧后产生二氧化硫并导致酸雨，用化学催化剂脱硫味道难闻又相当昂贵，这为大规模生物处理解决这一问题创造了条件，已经发现几株红球菌可将杂环上的硫通过氧化除去，面临的挑战是如何把这一反应规模化，以适应石油化工业对燃料的大规模需求。

红球菌菌株 I24 基因组测序正在进行，正在研究用它生产精细化工产品——光学构象纯 1-氨基-2-羟基茚满（1-amino-2-hydroxyindan)，该化学物质是茚地那韦（Indinavir，一种治疗人类免疫缺陷病毒的新药）的关键结构组成。红球菌菌株 I24 是少数可将茚（indene）氧化为顺-1,2-二氢二醇（*cis*-1,2-dihydrodiol）的菌株之一。用化学合成法可以把顺-1,2-二氢二醇转化为 1-氨基-2-羟基茚满[41]，关键问题是茚满二醇（indandiol）的立体化学纯度和氧化茚满（indan）产生的其他副产品氧化物。为手性（chiral）药物生产手性中间产物，是生物工程中竞争越来越激烈的领域[1]。随着对控制化合物立体特性（stereospecificity）的控制因素——酶的进一步了解，基因组学将会在这些领域发挥越来越大的作用。

药物生产一些起生物催化作用的最主要微生物几乎都是链霉菌属（*Streptomyces*）的成员，在生产天然产物和药用化合物，如抗生素、抗肿瘤因子和免疫抑制剂（immunosuppresant）方面，链霉菌在细菌各个属中排行第一。这些化合物的种类，包括聚酮、查尔酮和非核糖体多肽（nonribosomal peptide)，因此，正焦急期盼天蓝色链霉菌

菌株 A3 [2]的基因组测序的完成（表 2)[42]，该菌株 8.7 Mb 的巨大线性染色体，包含编码放线紫红素、土臭素和 coelichelin 的生物合成基因，所编码的基因数（7825）是已发现细菌中最多的，将会对未来天然产物的开发提供依据。

生物催化和生物降解的信息学

基因组序列数据只有与现有原核生物代谢活动的信息相互补充，才能最大地发挥它的作用，典型原核生物基因组的 75% 以上都编码蛋白，因此，基因注释在多数情况下涉及到把 DNA 序列翻译成催化某个生化反应或某一系列相关反应的蛋白质，越来越多的微生物反应正被互联网上的数据库归纳和收录，供大家共享。

代谢数据库非常专业化，往往集中于某一细菌菌株或广泛按代谢反应类型组建，例如，有共同中间产物的代谢反应，或按生物催化/生物降解来归类。EcoCyc 是专业代谢数据库，它专门描述大肠杆菌的代谢反应（表 4)。京都基因和基因组百科全书（Kyoto Encyclopedia of Genes and Genomes, KEGG)[43]是较广泛的数据库，它描述很多不同微生物中存在的代谢途径，大部分是中间产物的代谢，这些代谢途径分成各种类型，例如氨基酸代谢、核酸代谢或碳水化合物代谢（见第 6 章)。

表 4　微生物代谢代表性网上数据库，侧重生物催化和生物降解代谢

数据库	网址	覆盖范围
EcoCyc	http://ecocyc.org	大肠杆菌的代谢
KEGG	http://www.genome.ad.jp/kegg/	可点击的代谢图；广泛代谢
BSD	http://bsd.cme.msu.edu/bsd/	有重要生物降解价值的原核生物
UM-BBD	http://umbbd.ahc.umn.edu/	中间代谢之外的微生物代谢
UM-BBD 功能团 (UM-BBD Functional Groups)	http://umbbd.ahc.umn.edu/search/FuncGrps.html	酶催化特殊化学功能团的转化
UM-BBD 元素周期表 (UM-BBD Periodic Table)	http://umbbd.ahc.umn.edu/periodic/	微生物对化学元素的转化

明尼苏达大学生物催化/生物降解数据库（University of Minnesota Biocatalysis/Biodegradation Database, UM-BBD)[44]，正在努力按照生物降解和生物催化的微生物、基因、酶及其底物分类编纂数据。密歇根州立大学（Michigan State University）生物降解菌株数据库（Biodegradative Strain Database, BSD）（表 4）是 UM-BBD 互补数据库，它按照微生物菌株编纂数据。BSD 和 UM-BBD 都建立了交互链接，使用户无论首先搜寻微生物、酶、还是特定底物，都能得到全面生物降解方面的信息。

UM-BBD 构建者试图使该数据库能代表微生物生物催化反应的广度，无论这些反应正在被工业化利用，还是只描述自然界中微生物丰富多彩的生化反应。自然界中存在类型广泛的天然化合物，为土壤微生物的生长提供了潜在的底物，这表明自然界中微生物代谢类型的多样性。目前已知的化学物质有一千八百万种（18 million)，而且每天都

有新物质被合成，它们中的许多物质最终都可能被地球上某个角落的微生物所降解。也就是说，假定一种底物只进行一个反应，也就有上百万种反应正在地球上进行，这标志着代谢生物化学的发展前沿。在上世纪中，曾集中大量精力也只弄清了微生物中间代谢的大致轮廓，也就是大多数生化教科书里所介绍的那些内容，因此，从基因组测序项目中得到相当一部分未知基因，都有可能参与天然产物的分解或化学物质的合成[1]。

为了有助于基因组注释，UM-BBD 在官能团一节（表 4），列举了能被微生物产生的酶所作用的有机官能团，这些酶可以来自一种微生物，也可以来自一群协同作用的微生物。有一项新课题提供了关于微生物转化不同化学元素的代谢途径线索，除了生物体系中最常见的 24 种元素（碳、氢、氧、氮、磷、硫、钾、钠、镁、钙、硒、铁、锰、钒、钼、钴、镍、锌、硼、氯、溴、氟、碘、砷）外，微生物还能作用于、甚至还能改变很多元素的氧化态或化学种类（chemical speciation），这些反应能在生物治理中起重要作用，例如把铀还原，从而把它在土壤中的移动性降到最低（图 4）[45]。此外，微生物对矿物质的转化过程，对地球表面矿物质的沉积起重要作用，从而对这些元素在全球范围内的往复循环起重要作用。

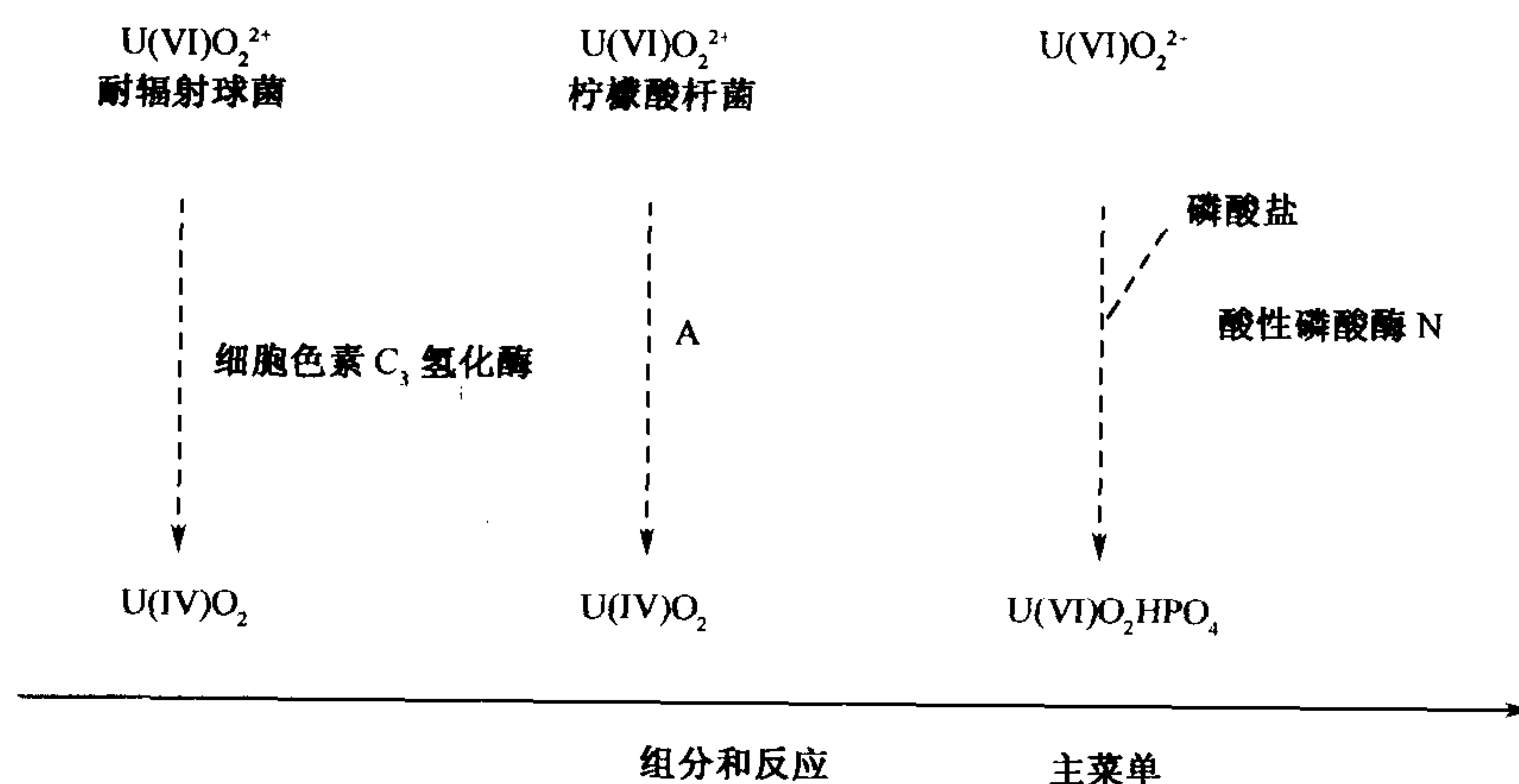


图 4　明尼苏达大学生物催化/生物降解数据库（UM-BBD）中细菌对铀的代谢图。

结语

微生物学的覆盖面很广，目前只能纯培养世界上不到 1% 的细菌。大自然进化的新种和代谢质粒的速度相当快，在这种不断变化情况下，微生物基因组测序仍在继续进行，并涵盖了原核生物更多种系。这些研究无疑揭示出微生物界丰富的基因组和生物催化反应，帮助了解生命，提供有工业价值的生物反应类型。生物催化反应及其作用酶的不断发现，提供了越来越多生物反应类型，从而提高了注释基因组的能力。因此，生物降解、生物催化以及基因组学这三个领域将不可避免地共同向前发展。

致谢

本人感谢 Steven Toeniskoetter 协助准备手稿。耐辐射异常球菌的电子显微照片由 Michael Daly 教授惠赠。

（许朝晖　译）

参考文献

1. Wackett LP, Hershberger CD. Biocatalysis and Biodegradation: Microbial Transformation of Organic Compounds. Washington, DC: American Society for Microbiology Press, 2001.
2. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 1998; 95:6578–6583.
3. D'Souza TM, Merritt CS, Reddy CA. Lignin-modifying enzymes of the white rot basidiomycete *Ganoderma lucidum*. Appl Environ Microbiol 1999; 65:5307–5313.
4. Kuan IC, Tien M. Stimulation of Mn peroxidase activity: a possible role for oxalate in lignin biodegradation. Proc Natl Acad Sci USA 1993; 90:1242–1246.
5. Nagasawa T, Yamada H. Bioconversion of nitriles to amides and acids. In: Abramowicz DA (ed). Biocatalysis. New York: Van Nostrand Reinhold, 1990, pp. 277–318.
6. Harbourne J. Ecological Biochemistry, 3rd ed. New York: Academic Press, 1988.
7. Vokounova M, Vacek O, Kunc F. Degradation of the herbicide bromoxynil in *Pseudomonas putida*. Folia Microbiol 1992; 37:122–127.
8. Blattner FR, Plunkett G, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997; 277:1453–1462.
9. Diaz E, Ferrandez A, Prieto MA, Garcia JL. Biodegradation of aromatic compounds by *Escherichia coli*. Microbiol Mol Biol Rev 2001; 65:523–569.
10. Krone UE, Thauer RK, Hogenkamp HP, Steinbach K. Reductive formation of carbon monoxide from CCl4 and FREONs 11, 12, and 13 catalyzed by corrinoids. Biochemistry 1991; 30: 2713–2719.
11. Ohren A, Gurevich P, Azachi M, Henis Y. Microbial degradation of pollutants at high salt concentrations. Biodegradation 1992; 3:387–398.
12. den Dooren de Jong LE. Bijdrage tot de kennis van het mineralisatieproces. Rotterdam, The Netherlands: Nijgh and van Ditmar, 1926.
13. Blumer M. Polycyclic aromatic compounds in nature. Sci Am 1976; 234:35–45.
14. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 1997; 390:249–256.
15. Torsvik V, Ovreas L. Microbial diversity and function in soil: from genes to ecosystems. Curr Opin Microbiol 2002; 5:240–245.
16. Lederberg J. Cell genetics and hereditary symbiosis. Physiol Rev 1952; 32:403–430.
17. Austen RA, Dunn NW. Isolation of mutants with altered metabolic control of the NAH plasmid-encoded catechol meta-cleavage pathway. *Aust J Biol Sci* 1977; 30:583–592.
18. Chaudhry GR, Huang GH. Isolation and characterization of a new plasmid from a *Flavobacterium* sp which carries the genes for degradation of 2,4-dichlorophenoxy-acetate. J Bacteriol 1988; 170:3897–3902.

19. Rheinwald JG, Chakrabarty AM, Gunsalus IC. A transmissible plasmid controlling camphor oxidation in *Pseudomonas putida*. *Proc Natl Acad Sci USA* 1973; 70:885–889.
20. Tralau T, Cook AM, Ruff J. Map of the IncP1β plasmid pTSA encoding the widespread genes (*tsa*) for *p*-toluenesulfonate degradation in *Comamonas testosteroni* T2. Appl Environ Microbiol 2001; 67:1508–1516.
21. Williams PA, Murray K. Metabolism of benzoate and the methylbenzoates by *Pseudomonas putida* (arvilla) mt-2: evidence for the existence of a TOL plasmid. J Bacteriol 1974; 1:416–423.
22. Wackett LP, Sadowsky MJ, Martinez B, Shapir N. Biodegradation of atrazine and related triazine compounds: from enzymes to field studies. Appl Microbiol Biotechnol 2002; 58:39–45.
23. Romine MF, Stillwell LC, Wong KK, et al. Complete sequence of 184-kilobase catabolic plasmid from *Sphingomonas aromaticivorans* F199. J Bacteriol 1999; 181:1585–1602.
24. Martinez B, Tomkins J, Wackett LP, Wing R, Sadowsky MJ. Complete nucleotide sequence and organization of the atrazine catabolic plasmid pADP-1 from *Pseudomonas* sp ADP. J Bacteriol 2001; 183:5684–5697.
25. Thorsted PB, Marcarteney DP, Akhtar P, et al. Complete sequence of the IncPβ plasmid R751: implications for the evolution and organisation of the IncP backbone. J Mol Biol 1998; 282: 969–990.
26. Raillard S-A, Krebber A, Chen Y, et al. Novel enzyme activities and functional plasticity revealed by recombining homologous enzymes. Chem Biol 2001; 125:1–9.
27. Riley RG, Zachara JM, Wobber FJ. Chemical contaminants on DOE lands, DOE/ER-0547T. Washington, DC: US Department of Energy, Office of Energy Research, Subsurface Science Program, 1992.
28. Johnson J. Hanford on fast forward. Chem Eng News, 2002; June 10:24–33.
29. Daly MJ. Engineering radiation-resistant bacteria for environmental biotechnology. Curr Opin Biotechnol 2000; 11:280–285.
30. White O, Eisen JA, Heidelberg JF, et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 1999; 286:1571–1577.
31. Karlin S, Mrazek J. Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. Proc Natl Acad Sci USA 2001; 98:5240–5245.
32. Lange CC, Wackett LP, Minton K, Daly M. Engineering a recombinant *Deinococcus radiodurans* for organopollutant degradation in radioactive mixed waste environments. Nature Biotech 1998; 16:929–933.
33. Brim H, McFarlan SC, Fredrickson JK, et al. Engineering *Deinococcus radiodurans* for metal remediation in radioactive mixed waste environments. Nature Biotechnol 2000; 15:85–90.
34. Silver S, Phung LT. Bacterial heavy metal resistance: new surprises. Annu Rev Microbiol 1996; 50 :753–789.
35. Fredrickson JK, Kostandarithes HM, Li SW, Plymale AE, Daly MJ. Reduction of Fe(III), Cr(VI), U(VI), and Tc(VII) by *Deinococcus radiodurans* R1. Appl Environ Microbiol 2000; 66:2006–2011.
36. Dixon B. Power Unseen: How Microbes Rule the World. Oxford, UK: Oxford University Press, 1984.
37. Kluyver AJ. Microbiology and industry. In: Kamp AF, La Riviere JWM, Verhoeven W (eds). A. J. Kluyver: His Life and Work. Amsterdam: North-Holland, 1957, pp. 165–185.
38. Glazer AN, Nakaido H. Microbial Biotechnology: Fundamentals of Applied Microbiology. New York: Freeman, 1995.
39. Lamed R, Setter E, Bayer EA. Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. J Bacteriol 1983; 156:828–836.

40. Gray KA, Pogrebinsky OS, Mrachko GT, Xi L, Monticello DJ, Squires CH. Molecular mechanisms of biocatalytic desulfurization of fossil fuels. Nature Biotechnol 1996; 14:1705–1709.
41. Treadway SL, Yanagimachi KS, Lankenau E, Lessard PA, Stephanopoulos G, Sinskey AJ. Isolation and characterization of indene bioconversion genes from *Rhodococcus* strain I24. Appl Microbiol Biotechnol 1999; 51:786–793.
42. Bentley SD, Chater KF, Cerdeno-Tarraga AM, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 2002; 417:141–147.
43. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 2002; 30:402–404.
44. Ellis LBM, Hershberger CD, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: microorganisms, genomics and prediction. Nucleic Acids Res 2000; 28: 377–379.
45. Lovley DR, Widman PK, Woodward JC, Phillips EJ. Reduction of uranium by cytochrome c3 of *Desulfovibrio vulgaris*. Appl Environ Microbiol 1993; 59:3572–3576.

25

酶的发现与微生物基因组学

Robert M Kelly and Keith R Shockley

引言

二十多年前，酶工业主要用狭窄种系发生范围的微生物集中生产少量的生物催化剂，这些酶应用范围有限，主要用于大规模淀粉加工和洗涤用的清洗剂[1~3]。如何发现新酶、如何大规模生产以及如何将它们合理用到现有的生物工艺中等存在诸多困难，从而限制了这一技术领域的发展。即使发现了具有生物催化剂潜在生理特征的一种微生物，从成千上万类似的生物分子中分离一种特殊酶，仍然具有相当大的挑战性。如果这种微生物（野生型或突变型）不具备用于大规模发酵产生丰富而大量酶的表型特征，那么这种生物催化剂要发挥它重要用途的可能性就相当低。简而言之，在 20 世纪 80 年代工业生物技术到来前夕，仅仅有少数几个工业先驱，能从隐藏在非典型特征微生物基因型内部，找出有价值的酶并实现商业化。

但是，今非昔比，自从第一个产酶重组生物体（经常涉及在天然寄主中进行表达基因的扩增；成功实现商业化开始，分子生物学彻底改变了酶发现的本质[4]。曾认为酶很少具有多种应用潜力，但是现在恰恰相反，根据已测序微生物基因组编码的信息推断，很多酶有不同的潜在应用价值，所面临的挑战是要为特定生物催化剂找到适合的技术应用领域。此外，定点突变、DNA 改组技术和定向进化等重组方法，已经用于创造能实际应用一些特定酶的工程产品[5,6]，同时，那些在关注生物催化剂的工艺学家们，正在探索有关那些珍贵酶类多样性的答案，并也在寻找能否采用生物途径代替或生产那些具有重要经济价值的化学和生物化学制剂。

许多工业用酶都传统地来源于天然环境或特殊生态位中分离的微生物，因为这些天然环境或特殊生态位中的生物过程与工业应用过程类似[4]，然而，传统酶发现的方法仅仅揭示了天然生物催化剂系列的一部分。rRNA 的小亚基（16s 或 18s）序列比较表明，地球上的大多数生命是微生物，然而，据估计自然界的微生物中 99% 以上还不能通过标准技术培养[7,8]。何况，已经确认的微生物还不到能培养微生物总数的 2%，仔细研究的微生物更是寥寥无几[9]。当前研究的微生物酶是否具有代表性？用现有微生物基因组序列资料可以对这一问题进行研究。

到目前为止，还不清楚从已测序微生物基因组中，能获得多少微生物多样性的信息，但可以肯定，在已测序的微生物基因组中，甚至在亲缘关系紧密的不同种或同一种内不同菌株之间，都存在着明显的遗传多样性[10]。直到现在，病原微生物一直是大多数基因组比较研究的主体[11~15]，但是，非病原微生物也开始很好地研究，例如，耐盐芽孢杆菌（*Bacillus halodurans*）中近三分之一的可读框与它同属的另一微生物——枯草芽孢杆菌（*Bacillus subtilis*）的可读框并没有明显的匹配[10]；再如，激烈火球菌

表 1　对微生物系统有用的生物信息学工具[a]

搜索工具	网　址	描　述
蛋白序列数据库		
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	基本局部联配搜索工具；用于序列的快速比较[34]；比较数据库中序列的快速途径，识别基因或病毒序列，并寻找一个感兴趣的序列与数据库序列的相似区域；由 BLAST 演变的程序有较高的敏感性，如蛋白序列一致性 BLAST (PSI-BLAST)或缺口 BLAST (Gapped-BLAST)
PROSITE	http://ca.expasy.org/prosite	通过与已知蛋白家族比较，阐明未知蛋白(从 cDNA 或基因组序列翻译得到的蛋白序列)的功能[53]
Pfam	http://pfam.wustl.edu/hmmsearch.shtml	多个蛋白域对比数据库，能评价和识别带有多个域的蛋白，检测蛋白序列间从头到尾的相似性[91]
Blocks	http://blocks.fhcrc.org	检测蛋白质局部区域相似性[92]
eMOTIF	http://motif.stanford.edu/emotif	判断并搜索蛋白质模体[93]
SMART	http://smart.embl-heideberg.de	简单的模块结构搜索工具；可以分析区域结构[94]
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS	保守蛋白质模体组的摘要[95, 96]
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	保守区域数据库；由蛋白保守区域的多个序列对比组成[97]
TOPITS	http://www.embl-heideberg.de/predictprotein/predictprotein.html	以预测为基础的穿线程序；用于推测蛋白质序列模体结构或功能[53]
基因组比较		
STRING	http://www.bork.embl-heideberg.de/STRING	搜索邻近基因再现的工具，在已公布的基因组序列中定位成簇重复出现的基因[98]；重复基因簇在不同基因组序列中出现，常常表明它们功能的相关性
COG	http://www.ncbi.nlm.nih.gov/COG	直系同源蛋白组集群，通过比较完整基因组中蛋白序列来决定；直系同源物通常具有相同功能[99, 100]
PEDANT	http://pedant.gsf.de/	高通量处理基因组数据，用广泛生物信息学方法给蛋白指定功能和结构类别[101]
AlignACE	http://arep.med.harvard.edu/mrnadata/mrnasoft.html	对比核酸保守元件；根据比较基因组学预测功能的相互作用[102, 103]
基因组信息代理	http://gib.genes.nig.ac.jp	可以获取和查阅任何已测序微生物基因组中感兴趣的区域[104]，并给出生物学释义

续表

搜索工具	网址	描述
代谢数据库		
EcoCyc	http://ecocyc.org	对大肠杆菌中所有已知代谢途径和信号传导途径进行注释，包括对机体基因组和生化机制的描述[65]
ENZYME	http://www.expasy.org/enzyme/	提供与感兴趣酶相关的系统命名、催化活性和辅助因子[105]
LIGAND	http://www.genome.ad.jp/ligand/	由三个主要区域组成；提供与有机体生物和化学特性相关联的信息；COMPOUND 提供代谢物和相关化学化合物信息；REACTION 收集了代谢反应；ENZYME 提供感兴趣蛋白所有已知酶反应[106]
KEGG	http://www.genome.ad.jp/kegg	京都基因和基因组百科全书；从基因组序列数据中提供功能信息[64]
WIT2	http://wit.mcs.anl.gov/WIT2/	What Is There 数据库；包含代谢途径信息，建立在序列比较和生化与表型数据的基础上[107]
其他有用的工具		
SignalP	http://www.cbs.dtu.dk/services/SignalP	在原核和真核系统神经网络基础上，识别信号肽和切割位点[108]
TMpred	http://www.ch.embnet.org/software/TMPRED_form.html	根据 TMbase 算法，预测蛋白质的跨膜区域和方向[109]
BRENDA	http://www.brenda.uni-koeln.de/	酶功能数据库[110]
ClustalW	http://www.ebi.ac.uk/clustalw/	DNA 或蛋白质多序列对比程序，以观察两个分子相似性或区别[111]
PSORT	http://psort.nibb.ac.jp/form.html	根据氨基酸序列数据，预测蛋白质分选信号或蛋白质在细胞中的定位[112]
TMHMM	http://www.cbs.dtu.dk/services/TMHMM	根据隐式马可夫模型，预测跨膜螺旋结构[113]
DAS	http://www.sbc.su.se/~miklos/DAS/	用"表面紧致排列"算法，根据氨基酸序列数据中预测任何镶嵌膜蛋白的跨膜区[114]
TIGRFAMs	http://www.tigr.org/TIGRFAMs/	根据隐式马可夫模型，确定蛋白质功能[115]

[a]对更广泛的生物信息学数据库列表见参考文献[116]和[117]。

(*Pyrococcus furiosus*) 的基因组[16]比同属的另一微生物——霍氏火球菌 (*Pyrococcus horikoshii*) 的基因组大 10%左右[17]。多数差异归咎于额外氨基酸生物合成途径和碳水化合物吸收途径，如纤维二糖、麦芽糖、海藻糖、昆布多糖和几丁质[16,18]。种内基因组序列比较的有限观察显示，目前对自然界生物催化剂成员的认识仅仅是窥见一斑。

很显然，一个个地研究基因和蛋白质的方法，正在被分子生物学和基因组学进步所带来的新信息和方法学的应用所替代[9,19,20]。当 20 世纪 90 年代中期完成的微生物基因组序列面世时，观察微生物种内全部酶库的组成才成为可能。现在，140 多种微生物基因组测序已经完成，至少还有 300 多个项目正在进行中[21]，通过微生物遗传信息而间接推断某些特定生物催化剂的存在才成为可能。

然而，由于微生物基因组中有一半甚至更多可读框，最初并没有特定的功能与之匹配，所以仍然有很多不确定性存在。例如，即使对研究最多的大肠杆菌，在 4288 个被释义的编码蛋白基因中，起初也有 38%没有确切功能[22]；再如，由极端嗜热酸的圆齿古生菌 (crenarcheaon)、硫磺矿硫化叶菌 (*Sulfolobus solfataricus*) P2 预测编码的 2977 个蛋白质中，约三分之一在其他已测序基因组中检测不到同源物[23]。在霍氏火球菌 (*P. horikoshii*) 基因组中，有 50%以上可读框的功能还无法通过数据库的相似性比较得以确认[24]。对微生物基因组中可读框的正确释义仍处在不断探索的阶段，当前主要采用体内、体外和计算机模拟 (*in silico*) 等多样化工具，目前，这个释义过程正像它已为许多问题提供了答案一样，也可能带来许多新问题，这已成为一个不可争议的事实。然而，微生物机体中与酶工业有关酶库的组成正变得越来越清晰。

随着用聚合酶链反应从基因组 DNA 中扩增有意义的基因，用于亚克隆并在适合寄主中超量表达，每一种微生物的基因组，都可能提供成千上万种具有重要意义的生物催化剂。通过应用各种各样生物信息学工具 (表 1)，可以进一步研究候选酶的特征，为有效缩短候选酶的名单，就像开发糖基水解酶那样[25]，通过计算机模拟酶结构及其催化特性，而且最好与以氨基酸序列为基础的分类系统相结合。

发掘超嗜热生物基因组，开发有用生物催化剂

因种种原因，最初测定的微生物基因组序列，主要集中在那些生存在对生物极端不利环境中的微生物[26]。超嗜热菌、嗜极端环境的微生物属古生菌域和细菌域，在 80℃ 或更高温度下最适合生长，由于它们的进化地位、小基因组以及产生稳定生物催化剂的特性，一些此类微生物的基因组序列已有报道 (表 2)。

尽管现有大量重组技术可以改善酶的特性 (见下文)，但是最好还是用具天然特性或生产特性最接近的酶。工业用生物催化剂极其受欢迎的特性就是热稳定性，这也是超嗜热微生物产生酶的本质特性。对温度稳定型生物催化剂感兴趣是因为：即使许多反应在不断提高的温度下进行，但大多数工业加工过程仍然使用来自嗜中温微生物的酶[4,27~29]。在较高温度下，有机化合物黏度降低和分散系数增大，会降低所需酶浓度并提高酶的转化[27,30,31]，温度升高也能使底物的生物利用率提高，同时可降低生物污染的风险[26,27,31]。另一优点是，超嗜热生物的酶经常对化学变性剂，如去污剂、促溶

表 2 超嗜热微生物基因组序列

名称	年份	描述	最适温度/℃	基因组大小/Mbp	可读框(ORF)	未知功能基因[a]	特有基因[a]	G+C/%	分离位置	参考文献
古细菌(Archaea)										
敏捷气热菌(*Aeropyrum pernix*)	1999	严格需氧的泉古细菌门	95	1.67	2694	523(19%)	1538(57%)	56	Kodakara	118
闪烁古细球菌(*Archaeoglobus fulgidus*)	1997	严格厌氧的古细球菌目,硫代谢	83	2.18	2436	1315(54%)	641(25%)	49	Vulcano	44
詹氏甲烷球菌(*Methanococcus jannaschii*)	1996	厌氧、营养缺陷和产甲烷的甲烷球菌目	85	1.66	1729	1076(62%)	525(30%)	31	东太平洋海丘	45
埃氏火球菌(*Pyrococcus abyssi*)	2001	厌氧火球菌目	98	1.77	1765	NR	NR	45	北斐济盆地	119
霍氏火球菌(*Pyrococcus horikoshii*)	1998	厌氧和专性异养的火球菌目	98	1.74	2061	859(42%)	453(22%)	42	冲绳岛海槽	17
激烈火球菌(*Pyrococcus furiosus*)	2002	厌氧的火球菌目,在有糖和肽时生长很好	98	1.91	2208	NR	NR	40	Vulcano	43
硫磺矿硫化叶菌(*Sulfolobus solfataricus*)	2001	需氧硫化叶菌目,低 pH 时生长最好	80	2.99	3032	577(22%)	743(25%)	NR	Pisciarelli Solfatara	23
需氧热棒菌(*Pyrobaculum aerophilum*)	2002	兼性需氧还原硝酸盐泉古细菌门	100	2.22	2587	NR	302(12%)	51	Maronti Beach	120
细菌(Bacteria)										
风产液菌(*Aquifex aeolicus*)	1998	微需氧的产液菌科;专性化能无机营养菌	95	1.55	1512	663(43%)	407(27%)	43	未报道	121
海栖热袍菌(*Thermotoga maritima*)	1999	厌氧热袍菌目,代谢简单和复杂的碳水化合物	80	1.86	1877	863(43%)	373(20%)	46	Vulcano	122

注:%G+C,鸟嘌呤和胞嘧啶的百分比;[a]指基因组序列公布时间;NR,未报道。

剂（chaotropic agents）和有机溶剂具有抵抗力，这使它们作为工业生物催化剂更有用[26,31~33]。因此，如果稳定性重要，一种现有超嗜热酶可以通过重组方法加以修饰以改善其催化特性，由于许多超嗜热生物的基因组序列已经完成，就很有可能找到一种具有特定生物催化特性的热稳定酶，要么满足特殊需要，要么经过修饰达到特定的要求。

超嗜热生物基因组中生物催化剂所有成分的检测

基因组序列数据中的可读框往往是通过与数据库，如基因库（Genbank，见第3章）中的可读框进行全序列比较（如，Basic Local Alignment Search Tool，BLAST[34]）来释义。借助专业化数据库（如参考文献[35]）和手册（如参考文献[36]），用类似方法也可以确定所选择机体中特定酶的详细目录，例如，表3列出了从激烈火球菌（*P. furiosus*）基因组序列中推断出所有已知（已分离和具有典型生物化学特征）或假定（如用表1所展示的生物信息学工具推断）的蛋白酶[37]。表1中蛋白酶同源物的标准是：蛋白质50%以上的区域在氨基酸水平有30%以上的序列一致性，这个指定的标准可视具体情况而改变。表3展示出在一个指定机体内和机体之间蛋白酶的生物多样性，正如期望的那样，尽管这三个火球菌表现出明显的差异，它们仍然具有非常相似编码蛋白酶的基因。在有些情况下，列举的超嗜热生物中没有激烈火球菌蛋白酶的同源物，而在另一些情况下，似乎有同源物，但分子质量有明显差别。例如，一个推断的蛋白酶（带一个信号肽，PF1905），三个氨基肽酶（带信号肽、PF2059、PF2063和PF2065），一个推断的胞内细菌素/蛋白酶（PF1191）和一个转膜蛋白酶热溶素（pyrolysin）（PF0287）似乎是激烈火球菌特有的，而一种胞内邻-唾液酸糖蛋白内肽酶（PF0172）、一种脯氨酸二肽酶（PF1343）和一种甲硫氨酸二肽酶（PF0541）在所有超嗜热生物基因组序列中广泛存在（表3）。

最早研究超嗜热生物的酶中，有一些是糖基水解酶[38,39]，这些酶之所以引人注目，主要是它们在淀粉加工工业中的意义，以及它们对在葡聚糖培养基中超嗜热异养生长的重要生理作用[40,41]。超嗜热生物基因组序列揭示，它们代谢碳水化合物的水解酶有显著差异[42]。尽管超嗜热激烈火球菌的基因组序列[43]中存在许多降解葡聚糖的酶（表4），但是，另一种超嗜热古生菌——闪烁古细球菌（*Archacoglobus fulgidus*）的基因组序列中[44]明显缺少这种酶，甚至在三种火球菌中，糖基水解酶成分也有变化，尤其是能降解昆布多糖[40]和几丁质[44a]的酶。事实上，激烈火球菌有一条利用几丁质的途径，包括几丁质脱乙酰酶和糖氨基化酶（glucoaminidase），而这两种酶最初都未能从基因组释义中推导出来[44a]。另外，詹氏甲烷球菌（*Methanococcus jannaschii*）的基因组序列[45]，揭示有几种糖苷酶存在，其中一种（MJ1601）是第15族系葡糖淀粉酶。尽管在对这些生物进一步研究时可能会出现一些意外的结果，由于超嗜热生物基因组的数量较少，很难评价酶（如糖苷酶）的多样性。

在某种水平上，基因组序列的释义为特定微生物体中，实际和推断的酶目录提供了依据，然而，基因组序列释义本身对生物催化剂的鉴定也是一大挑战，因为有时的结果可能是误导[20,46]，例如，尽管用序列比对能将两种推断的酶归类为相关，并在序列释义时也如是反映，但是具有相似底物结合域的这两种酶，可能具有不同的催化域（见第3

表 3 激烈火球菌的蛋白酶

	基因座	S[a]	Nuc.	a..a.	Ph	Pa	Mj	Af	Pae	Ap	Ss	Tm	Aa
ATP-依赖型蛋白酶													
蛋白酶体,β 亚单位(PsmB-1)	PF1404	N	621	206	x	x	x	x	x	x	x		
蛋白酶体,β 亚单位(PsmB-2)	PF0159	N	599	199	x	x	x	x	x	x	x		
ATP-依赖型调节亚单位(PAN)	PF0115	N	1199	399	x	x	x	x		x	x		
ATP-依赖型 LA(Lon)	PF0467	N	3140	1046	x	x		x					
蛋白酶体,α-亚单位(PsmA)	PF1571	N	798	265	x	x	x	x	x	x	x		
ATP-不依赖型蛋白酶													
枯草芽孢杆菌蛋白酶样蛋白酶	PF0688	N	593	197		x			x	x[b]			
胞内蛋白酶 I(PfpI)	PF1719	N	582	193	x	x	x	x	x	x	x		x
假定细胞周质丝氨酸蛋白酶	PF0240	Y	842	280	x	x	x	x		x		x	x
金属蛋白酶	PF0392	N	1253	417	x	x		x		x			
碱性丝氨酸蛋白酶	PF1670	Y	1992	663				x	x	x			
金属蛋白酶	PF0167	Y	1133	377	x	x	x	x		x			
假定细菌素/蛋白酶	PF1191	N	785	261									
热解素	PF0287	Y	4238	1412					x[b]				
氢化酶成熟蛋白酶 (hyc I)	PF0617	N	485	161	x	x	x	x					
假设蛋白	PF0760	N	1022	340	x						x	x	
蛋白酶 IV	PF1583	Y	990	329	x	x	x						x
假定蛋白酶	PF1905	Y	1332	443				x[b]					
金属蛋白酶	PF0457	N	629	209		x	x		x		x		x
肽酶													
乙酰鸟氨酸脱乙酰基酶(ArgE)/肽酶	PF1185	N	1061	353	x	x			x				
HtpX 热激蛋白	PF1135	N	875	291	x	x	x	x		x			x
脯氨酸二肽酶相关蛋白	PF0702	N	521	173	x	x						x	
内-1,4-β-葡聚糖酶(ytoP)相似蛋白	PF1861	N	1040	346	x	x	x	x	x	x		x	
D-氨肽酶	PF1924	N	1098	365	x	x							

续表

	基因座	S[a]	Nuc.	a..a.	Ph	Pa	Mj	Af	Pae	Ap	Ss	Tm	Aa
信号序列肽酶 I, SEC 11	PF0313	N	264	87	x	x			x	x			
假定蛋白	PF0669	N	932	310	x	x							x[b]
邻-唾液酸糖蛋白内肽酶 (gcp-2)	PF0473	N	680	226	x	x		x	x	x	x		
邻-唾液酸糖蛋白内肽酶 (gcp-1)	PF0172	N	974	324	x	x	x	x	x	x	x	x	x
琥珀酰-二氨基庚二酸脱琥珀酰酶/肽酶	PF2048	N	1343	447	x	x				x			x[b]
焦谷氨酰肽酶 I	PF1299	N	654	217	x	x					x		
XAA-Pro 二肽酶 (脯氨酸二肽酶)	PF1343	N	1047	348	x	x	x	x	x	x	x	x	x
酰基氨酰肽酶相似蛋白(酰基氨基酸释放酶同系物)	PF0318	N	1862	620	x	x			x	x			
polyl 内肽酶 (polyl endopeptidase)	PF0825	N	1863	620	x	x							
内切葡聚糖酶(CelM)/氨肽酶	PF1547	N	1046	348	x	x	x	x	x	x	x[b]	x	
甲硫氨酸氨肽酶(MAP)(Pep M)	PF0541	N	887	295	x	x	x	x	x	x	x	x	x
推断的脯氨酸二肽酶	PF0747	N	1076	358	x	x	x						
热激蛋白 X	PF1597	Y	800	266	x	x							
羧肽酶 I	PF0456	N	1499	499	x	x			x	x	x		
内切葡聚糖酶/肽酶	PF0369	N	998	332	x	x	x	x	x	x		x	
推断的氨肽酶	PF2059	Y	1704	567									
推断的氨肽酶	PF2063	Y	1755	584									
推断的氨肽酶	PF2065	Y	1776	591									
膜二肽酶	PF0874	HN	1140	379	x	x				x			

注:如果与数据库中蛋白相比,某蛋白质 50%以上氨基酸的一致性大于 30%,则认为该蛋白质在数据库中存在。表中数据为激烈火球菌(菌株 DSM3638)与其他菌的比较。Ph, 霍氏火球菌 OT3;Pa, 埃氏火球菌 GE5; Mj, 詹氏甲烷球菌 DSM2661; Af, 闪烁古细球菌 VC-16; Pae, 需氧热棒菌 IM2; Ap, 敏捷气热菌 KI; Ss, 硫磺矿硫化叶菌 P2; Tm, 海栖热袍菌 MSB8; Aa, 风产液菌 VF5。[a] S 根据 SignalP 预测信号肽的有无(Y, 有;N, 无)[108]; [b]表示预测蛋白的氨基酸长度与从激烈火球菌中预测蛋白的长度有显著区别。

表 4 激烈火球菌糖苷酶一览表(根据基因组序列资料)

基因座	注释	参考文献	活力	S[a]	Ph	Pa	Mj	Af	Pae	Ap	Ss	Tm	Aa
	纤维素酶												
PF0073	β糖苷酶	123,124	Cel1A		x								
PF0442	β糖苷酶	125	Cel1B						x				
PF0854	内切-1,4-β-葡聚糖酶	126	Cel12	Yes								x	
	甘露糖苷酶												
PF1208	β甘露糖苷酶	127	Man1		x	x							
	昆布多糖酶												
PF0076	内切-1,3-β-葡聚糖酶	123,128	Lam16	Yes								x	
	几丁质酶												
PF1234	几丁质酶	44a, 129	Chi18A	Yes									
PF1233	几丁质酶	44a, 129	Chi18B										
	淀粉酶/支链淀粉酶												
PF0477	α淀粉酶	130,131	Amy13	Yes			x					x	
PF0272	α淀粉酶	132	Amy57A		x	x	x		x				x
PF1935	淀粉支链淀粉酶	133	Amy57B			x			x		x		
PF0478	α淀粉酶[b]		Amy13								x		x
PF1939	新支链淀粉酶		Pul13						x		x	x	
	半乳糖苷酶												
PF0444	α半乳糖苷酶		Gal57		x								
PF0356	β半乳糖苷酶[b]		Gal1			x					x		
PF0636	β半乳糖苷酶前体[b]		Gal35		x	x							

[a] S-根据 SignalP 预测信号肽的有无(Y, 有;N, 无)[108]; [b]推测蛋白;黑体表示已研究清楚的酶。

章)。当根据全序列进行简单同源物搜索时，可能检测不到酶的超家族中不太明显的相关性，也不能识别带有同一功能的非直系同源基因[4,47,48]。当数据库中特定可读框被错认为具有某功能时，这种错误指定会在后续报道序列中累积，因此，仅仅凭借简单的氨基酸序列同源分析，不足以确定基因组序列中缺乏某种酶或排除某种酶具有多种功能的可能性。对细胞代谢的不完全理解也会出现一些问题，例如，微生物编码色氨酸生物合成途径中，一些酶的基因在霍氏火球菌（*P. horikoshii*）基因组中缺乏[24]，尽管该菌需要色氨酸维持细胞活力和生长，但还不清楚它是色氨酸营养缺陷型，还是存在一条完整包含未知成分的合成途径。

当单独用 BLAST 搜索工具不足以阐明基因功能时，可用其他生物信息学方法补充分析。例如，已报道嗜中温生物的酯酶和脂肪酶，可酶促分解 2-芳基丙酸酯的外消旋混合物（如那些用于非类固醇的抗炎症药物）[49]，将 BLAST 搜索与蛋白质结构模体分析（threading one-dimensional predictions into three-dimensional structures, TOPITS[50]）相结合，鉴定出硫磺矿硫化液菌（*S. solfataricus*）P1（SsoEST1）基因组中的一种羧酸酯酶[51]。已经证实，尽管测试温度比最适温度低 50℃ 多度，该酶比其他嗜中温候选酶能更有效地分解萘普生甲酯（naproxen methyl esters）[52]，单独 BLAST 搜索只给出了在其他温度稳定型的酯酶/脂肪酶，但与其他几种生物信息学工具联用，便发现了最有希望的候选物。

搜索数据库，如 PROSITE[53]来寻找短序列模式或模体，以鉴定预测蛋白质的功能域，能剔除全序列对比带来的问题[54]。例如，当 BLAST 比较没有任何提示性结果时，另一个 IDENTIFY[55]数据库也可能识别出蛋白质超家族。在酵母基因组首次公布时，有 833 个可读框没有指定功能，但根据 IDENTIFY 算法，有 172 种未知蛋白质相继被赋予假定的功能[55]。

其他方法也能应用，利用穿针引线法（threading）或从头折叠法（*ab initio* folding)，能根据序列信息预测蛋白质三级结构，然后通过分析蛋白质活性位点的程序——模糊功能形式（fuzzy functional form）进一步筛选三级结构[47,54]。酵母基因组中，有的基因功能不能通过 BLAST 搜索或局部序列比对预测，借助这种方法确认了酵母基因组中，谷氧还蛋白/硫氧还蛋白二硫键氧化还原酶家族中的 2 种蛋白的功能[54]。另一种方法建立在预测与实验数据相结合的基础上，研究了酿酒酵母（*Saccharomyces cerevisiae*）6217 种蛋白质之间的进化相关性、信使 RNA 表达模式相关性和区域融合模式，给 2557 种未知酵母蛋白中一半以上的蛋白指定了功能[56]。

比较不同微生物基因组也能发现有应用前途的生物催化剂，例如，跨越物种界线的保守基因簇，有助于确定有同源功能的蛋白质，或表明某种必需功能的存在。恶性疟原虫（*Plasmodium falciparum*）合成必需的类异戊二烯，所利用的是在植物体内普遍存在而在动物体内没有的一条生物合成途径，以该途径为靶点，发明了可作为特殊抗疟疾的药物[48,57]。有时，水平基因转移使研究单基因进化相关性变得更复杂，然而，通过寻找区域性差别，如细菌染色体中碱基组成的差别，可用全基因组的序列信息，更快地识别发生水平转移的基因[48]。

功能基因组学与酶的发现

识别的可读框等基因组信息，便对一个生物体的全部基因有一定认识，仅此还是无法知道生物体怎样利用这些遗传信息完成它的生物学使命，只有掌握了生物体是如何运作的知识，才能有目的地利用特定酶。遗传控制就是决定一个基因是处于活性状态还是失活状态的过程[58,59]。活性状态基因是那些正在转录、转录物正在翻译或翻译产物正在有效行使其功能。功能基因组学既包括转录分析又包括蛋白质组学方法，利用基因表达数据，系统地大规模诱变和蛋白质相互作用图谱阐明基因的功能[47,60]。在细菌和真核生物（可能还有古生菌）中，大多数基因调节经常控制在转录起始水平[48,61,62]，因此，对基因转录起始调控机制的理解是发现有意义酶的一条很好的途径，也是了解基本生命过程所必需的。

正如上所述，基因组序列比较是向阐明基因组信息编码某一蛋白在代谢中的角色迈出的第一步，但是，生物信息学预测必须经过转录分析和生物化学分析进一步确认。例如，海栖热袍菌（*Thermotoga maritima*）是能在含一系列 α 和 β 糖苷键化合物中生长的超嗜热细菌[63]，它产生几种糖基水解酶，包括内切葡聚糖酶（Cel5A）和甘露聚糖酶（Man5）。通过 BLAST 搜索将这两种酶与 Genbank 数据库中的蛋白比较后发现[42]，甘露聚糖酶（Man5）与嗜热脂肪芽孢杆菌（*Bacillus stearothemophilus*）中的 β 甘露聚糖酶（ManF）最相似，氨基酸序列一致性为 46%，而内切葡聚糖酶（Cel5A）与溶胞梭状芽孢杆菌（*Clostridium celluloyticum*）中第 5 族系的内切葡聚糖酶（CelD）有最高的相似性，氨基酸序列一致性为 38%。Northern 杂交和 cDNA 芯片实验证明，当海栖热袍菌生长在角豆半乳甘露聚糖、魔芋葡甘露聚糖和少量羧甲基纤维（CMC）上时，甘露聚糖酶基因（*man*5）可诱导表达，而内切葡聚糖酶基因（*cel*5*A*）仅在有魔芋葡甘露聚糖时被诱导表达。

为了进一步研究这个意外结果，以多种聚糖为底物检测甘露聚糖酶（Man5）和内切葡聚糖酶（Cel5A）的重组酶的活性，甘露聚糖酶（Man5）仅在甘露糖的聚糖培养物上有活力，而内切葡聚糖酶（Cel5A）则在葡聚糖、木聚糖和甘露聚糖上均有活性。有趣的是，在以半乳甘露聚糖为底物时，Cel5A 酶的活力与 Man5 酶的活力相当，而在葡甘露聚糖为底物时，Cel5A 酶的活力明显高于 Man5 酶的活力，这个结果与单独凭基因序列比较预测的结果差别很大。对这两个酶的进一步研究发现，Man5 酶含有一个信号肽，而 Cel5A 酶不含信号肽。总之，这些结果表明，Cel5A 酶（及相关的 Cel5B 酶）最基本的生理作用是：在胞外的糖苷酶（如 Man5）完成水解作用后，Cel5A 降解转运到细胞内的葡甘露聚糖的寡聚糖；尽管在基因组中编码 Cel5A 和 Man5 酶的基因并不相邻（图 1），但它们的功能却密切相关，Cel5A 酶的功能是根据酶的生物化学特性及其在各种底物上生长天然菌株的基因调节模式而确定的。

对已测序基因组编码酶的生物学功能，有时可以用数据库序列资料与研究得比较清楚的蛋白质比较，或用生物信息学工具，包括京都基因和基因组百科全书（Kyoto Encyclopedia of Genes and Genomes，KEGG）[64]和 EcoCyc[65]来确定，它们都在生物学途径和分子组装的基础上，详尽地展现了基因组信息。另一种办法是，对两个或多个不同样

本进行表达转录物差异性鉴定和定量，然后直接从表达分析中得到有用基因功能方面的数据。

对表达基因进行差异性鉴定和定量的研究方法已经取得很大进展，尤其是 cDNA 芯片技术[66~73]，它首次用于同时监测 45 种不同的拟南芥（*Arabidopsis*）基因的表达[74]，自此以后，DNA 芯片已经用于多种生物体大规模基因表达模式的研究，包括细菌[66~73]、古生菌[75]、酵母[76~78]、植物[79~80]、果蝇[81]、小鼠[82]和人类[83~85]。芯片技术给生长在特定底物上的微生物，提供了鉴定差异表达基因的机制，对芯片的修饰具有重要意义，例如，应用特定序列 cDNA 芯片发现，当海栖热袍菌（*T. maritime*）在含甘露聚糖化合物的培养基上生长时，编码 Cel5A 和 Man5 的基因是共调节的[86]。随着芯片技术的不断改进和价格的便宜，应用环境芯片来跟踪微生物聚生体（consortia）中的基因表达模式，可能成为发现生物催化剂的一种新途径。

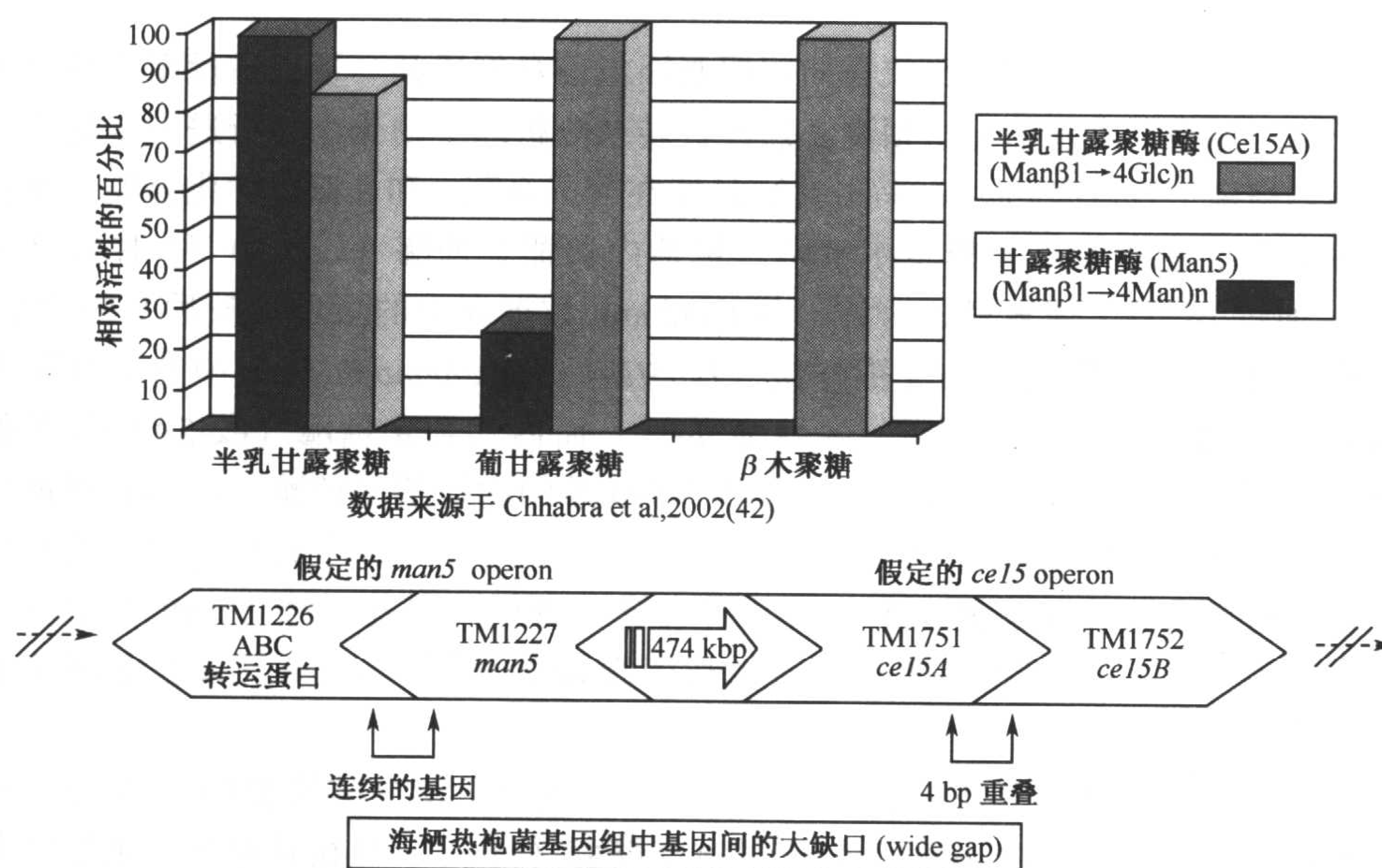


图 1 海栖热袍菌中的甘露聚糖酶和半乳甘露聚糖酶。图中展现了每个酶对半乳甘露聚糖酶、葡甘露聚糖酶和 β-木聚糖的酶活力，以及编码酶的基因在基因组中的位置[42]。

通过基因组扫描、功能筛选和改进设计生物催化剂

随着高通量筛选技术的逐年改进[87,88]，可以用重组方法改进生物催化剂，生产具有特殊功能的优质酶。直接分子进化技术及相关方法能够产生现有酶的有用变异体，使最终生物催化剂比天然的效果更好（见第 24 章），这种方法由多轮次诱变和筛选组成，然后对选择的变异体进行扩增[4,5,89]。随着对基因释义越来越完善，进化技术起战略性起点作用筛选生物分子的过程也越来越完善。

近来有个直接分子进化的例子是，在激烈火球菌（*P. furiosus*）中与功能不相关的一个 DNA 片段产生了氨苄青霉素抗性[89]，该突变体酶使细菌对靶向细胞壁合成的其他

药物也具有抵抗力，其作用机制尚不清楚。激烈火球菌是细胞壁成分不含肽聚糖的超嗜热古生菌，对一般抗细胞壁合成的抗生素不敏感，包括氨苄青霉素类的 β 内酰胺抗生素。尽管如此，还是筛选了激烈火球菌 DNA 片段表达文库，得到在大肠杆菌中有氨苄抗性（amp^R）长 1.2kb 的 DNA 片段（含编码 266 个氨基酸的一个可读框）[89]，然后对该片段进行了 50 次随机引入突变和 DNA 重组的直接进化。最终 DNA 片段在实验中含有 2 个共进化遗传区域，一个是氨苄抗性必需的，另一个能增强这种抗性。这个实验说明，从基因组序列中选择一些基因，使它们进化出能与其自身角色不相关、有实际应用价值的功能。

微生物基因组学：未来酶发现的方向

现在谈论在基因组学的基础上，如何有效地发现酶的方法还为时过早，除了利用生物信息学工具寻找感兴趣酶的同源物外，那些能产生重要酶或代谢途径的生理系统也极其有趣，通过差异表达实验可以推测这些生理系统，可以用全基因组芯片或特定序列芯片，研究微生物对环境或营养的改变所做出遗传学的应答反应。例如，激烈火球菌像其他超嗜热生物一样，缺乏磷酸转移酶系统，靠腺苷-三磷酸结合框透性酶吸收碳水化合物（见图 2）。糖苷酶与转运蛋白的偶联机制，可用来追踪细胞对各种碳水化合物的应答，从而给那些编码水解特定底物酶的基因提供释义线索。因此，通过对特殊碳水化合

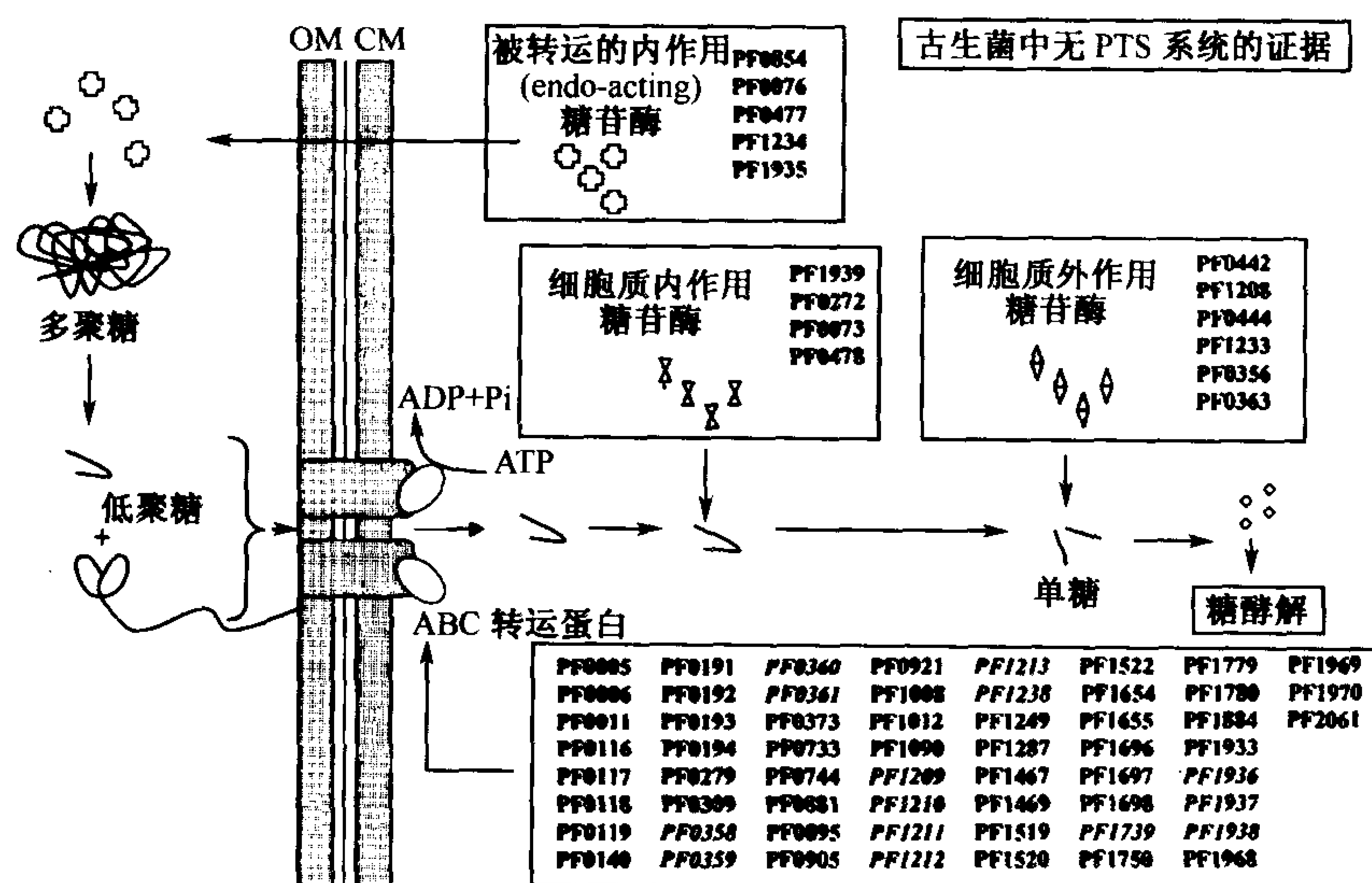


图 2 激烈火球菌中的糖代谢。包括所有由基因组序列信息和在线 CAZY 数据库（http://afmb.cnrs-mrs.fr/～cazy/CAZY/，按参考文献 25 的家族系统分类糖苷酶）识别的胞内和胞外糖苷酶及（ABC）转运蛋白。* 表示来自激烈火球菌基因组的已知和假定的 ABC 转运蛋白，包括那些在注释中划分到涉及糖运输或碳水化合物转运的蛋白家族和二肽/寡肽运输（Opp）家族[90]成员。斜体字表示激烈火球菌基因组中特征明确的转运蛋白或位于已知或假定的糖苷酶附近的转运蛋白。OM，外膜；CM，细胞膜；PTS，磷酸转移酶系统；Pi 自由磷酸。

物的差异表达分析，可以发现参与多糖各个水解阶段的新酶，如果对机体生理模式有足够认识，那么类似的方法也适于其他类型的酶。

自从酶发现的早期到现在，已经发生了很大变化。微生物基因组无疑会激发创造性地开发以前被埋没的生物催化剂，通过将最新发展的高通量筛选方法、直接进化方法和生物催化剂生产方法与生物信息学工具相结合，微生物基因组可以充分应用到与重要生物转化相关的重大技术进步中。

致谢：

本项目部分工作得到国家科学基金生物技术项目和美国能源部能量生物科学项目的资助。

（刘明秋 译）

参考文献

1. Dordick JS. The general uses of biocatalysts. In: Dordick JS (ed). Biocatalysts for Industry. New York: Plenum Press, 1991, pp. 1–19.
2. Neidleman SL. Historical perspective on the industrial uses of biocatalysts. In: Dordick JS (ed). Biocatalysts for Industry. New York: Plenum Press, 1991, pp. 21–33.
3. Uhlig H. Industrial Enzymes and Their Applications. Linsmaier-Bednar (EM (trans). New York: Wiley, 1998.
4. Marrs B, Delagrave S, Murphy D. Novel approaches for discovering industrial enzymes. Curr Opin Microbiol 1999; 2:241–245.
5. Arnold FH, Volkov AA. Directed evolution of biocatalysts. Curr Opin Chem Biol 1999; 3: 54–59.
6. Crameri A, Raillard SA, Bermudez E, Stemmer WP. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 1998; 391:288–291.
7. Aaman RI, Ludwig W, Schleifer K-H. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol Rev 1995; 59:143–169.
8. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 1998; 180:4765–4774.
9. Bull AT, Ward AC, Goodfellow M. Search and discovery strategies for biotechnology: the paradigm shift. Microbiol Mol Biol Rev 2000; 64:573–606.
10. Boucher Y, Nesbo CL, Doolittle WF. Microbial genomes: dealing with diversity. Curr Opin Microbiol 2001; 4:285–289.
11. Stephens RS, Kalman S, Lammel C, et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 1998; 282:754–759.
12. Alm RA, Ling LS, Moir DT, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 1999; 397:176–180.
13. Alm RA, Trust TJ. Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. J Mol Med 1999; 77:834–846.
14. Kalman S, Mitchell W, Marathe R, et al. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet 1999; 21:385–389.

15. Read TD, Brunham RC, Shen C, et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 2000; 28:1397–1406.
16. Robb FT, Maeder DL, Brown JR, et al. Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. Methods Enzymol 2001; 330:134–157.
17. Kawarabayasi Y, Sawada M, Horikawa H, et al. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement). DNA Res 1998; 5:147–155.
18. Lecompte O, Ripp R, Puzos-Barbe V, et al. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res 2001; 11:981–993.
19. Dean PM, Zanders ED, Bailey DS. Industrial-scale, genomics-based drug design and discovery. Trends Biotechnol 2001; 19:288–292.
20. Tang CM, Moxon ER. The impact of microbial genomics on antimicrobial drug development. Annu Rev Genomics Hum Genet 2001; 2:259–269.
21. The Institute for Genomic Research. Completed genomes available at: http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl; genomes in progress available at: http://www.tigr.org/tdb/mdb/mdbinprogress.html. Accessed November 24, 2003.
22. Blattner FR, Plunkett G, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. Science 1997; 277:1453–1462.
23. She Q, Singh RK, Confalonieri F, et al. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci USA 2001; 98:7835–7840.
24. Kawarabayasi Y. Genome of *Pyrococcus horikoshii* OT3. Methods Enzymol 2001; 330:124–134.
25. Henrissat B, Teeri TT, Warren RA. A scheme for designating enzymes that hydrolyse the polysaccharides in the cell walls of plants. FEBS Lett 1998; 425:352–354.
26. Adams MW, Perler FB, Kelly RM. Extremozymes: expanding the limits of biocatalysis. Biotechnology 1995; 13:662–668.
27. Zamost BL, Nielsen HK, Starnes RL. Thermostable enzymes for industrial applications. J Indust Microbiol 1991; 8:71–82.
28. Stetter KO. Hyperthermophiles: isolation, classification, and properties. In: Horikoshi K, Grant WD (eds). Extremophiles: Microbial Life in Extreme Environments. New York: Wiley-Liss, 1998, pp. 1–24.
29. Demirjian DC, Moris-Varas F, Cassidy CS. Enzymes from extremophiles. Curr Opin Chem Biol 2001; 5:144–151.
30. Kalisz MH. Microbial proteinases. Adv Biochem Eng Biotechnol 1988; 36:17–55.
31. Niehaus F, Bertoldo C, Kahler M, Antranikian G. Extremophiles as a source of novel enzymes for industrial application. Appl Microbiol Biotechnol 1999; 51:711–729.
32. von der Osten C, Branner S, Hastrup S, et al. Protein engineering of subtilisins to improve stability in detergent formulations. J Biotechnol 1993; 28:55–68.
33. Cowan DA. Protein stability at high temperatures. Essays Biochem 1995; 29:193–207.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403–410.
35. Barrett AJ, Rawlings ND, O'Brien EA. The MEROPS database as a protease information system. J Struct Biol 2001; 134:95–102.
36. Barrett AJ, Rawlings ND, Woessner JF. Handbook of Proteolytic Enzymes. London: Academic, 1998.
37. Ward DE, Shockley KR, Chang LS, et al. Proteolysis in hyperthermophilic microorganisms. Archaea 2002; 1:63–74.

38. Costantino HR, Brown SH, Kelly RM. Purification and characterization of an alpha-glucosidase from a hyperthermophilic archaebacterium, *Pyrococcus furiosus*, exhibiting a temperature optimum of 105 to 115 degrees C. J Bacteriol 1990; 172:3654–3660.
39. Brown SH. Saccharidases from high-temperature bacteria: physiological and enzymological studies. PhD thesis, Johns Hopkins University, Baltimore, MD, 1992.
40. Bauer MW, Driskill LE, Kelly RM. Glycosyl hydrolases from hyperthermophilic microorganisms. Curr Opin Biotechnol 1998; 9:141–145.
41. Driskill LE, Kusy K, Bauer MW, Kelly RM. Relationship between glycosyl hydrolase inventory and growth physiology of the hyperthermophile *Pyrococcus furiosus* on carbohydrate-based media. Appl Environ Microbiol 1999; 65:893–897.
42. Chhabra SR, Shockley KR, Ward DE, Kelly RM. Regulation of endo-acting glycosyl hydrolases in the hyperthermophilic bacterium *Thermotoga maritima* grown on glucan- and mannan-based polysaccharides. Appl Environ Microbiol 2002; 68:545–554.
43. Weiss RB. Direct Submission: *Pyrococcus furiosus* genomic sequence. Salt Lake City, UT: Human Genetics, University of Utah, 2002.
44. Klenk HP, Clayton RA, Tomb JF, et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 1997; 390:364–370.
44a. Gao J, Bauer MW, Shockley KR, Pysz MA, Kelly RM. Growth of hyperthermophilic archaeon *Pyrococcus furiosus* on chitin involves two family 18 chitinases. Appl Environ Microbiol 2003; 69:3119–3128.
45. Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 1996; 273:1058–1073.
46. Pennisi E. Keeping genome databases clean and up to date. Science 1999; 286:447–450.
47. Pallen MJ. Microbial genomes. Mol Microbiol 1999; 32:907–912.
48. Sassetti C, Rubin EJ. Genomic analyses of microbial virulence. Curr Opin Microbiol 2002; 5: 27–32.
49. Sehgal D. Effect of reaction environment on biocatalysis and enantioselectivity of hyperthermophilic esterases. PhD thesis, North Carolina State University, Raleigh, NC, 2002.
50. Rost B. TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc Int Conf Intell Syst Mol Biol 1995; 3:314–321.
51. Sehgal AC, Callen W, Mathur EJ, Short JM, Kelly RM. Carboxylesterase from *Sulfolobus solfataricus* P1. Methods Enzymol 2001; 330:461–471.
52. Sehgal AC, Kelly RM. Enantiomeric resolution of 2-aryl propionic esters with hyperthermophilic and mesophilic esterases: contrasting thermodynamic mechanisms. J Am Chem Soc 2002; 124:8190–8191.
53. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res 1999; 27:215–219.
54. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998; 281:949–968.
55. Nevill-Manning CG, Wu TD, Brutlag DL. Highly specific protein sequence motifs for genome analysis. Proc Natl Acad Sci USA 1998; 95:5865–5871.
56. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999; 402:83–86.
57. Jomaa H, Wiesner J, Sanderbrand S, et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 1999; 285:1573–1576.
58. Kornberg RD. Eukaryotic transcriptional control. Trends Biochem Sci 2000; 24:M46–M49.

59. Klug WS, Cummings MR. Genetics. Upper Saddle River, NJ: Prentice-Hall, 2000.
60. Kotra LP, Vakulenko S, Mobashery S. From genes to sequences to antibiotics: prospects for future developments from microbial genomics. Microbes Infect 2000; 2:651–658.
61. Neidhardt FC, Ingraham JL, Schaechter M. Physiology of the Bacterial Cell: A Molecular Approach. Sunderland, MA: Sinauer Associates, 1990.
62. Thomm M. Archaeal transcription factors and their role in transcription initiation. FEMS Microbiol Rev 1996; 18:159–171.
63. Huber R, Langworthy TA, Konig H, et al. *Thermotoga maritima* sp nov represent a new genus of unique extremely thermophilic eubacteria growing up to 90°C. Arch Microbiol 1986; 144: 324–333.
64. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res 2002; 30:42–46.
65. Karp PD, Riley M, Saier M, et al. The EcoCyc Database. Nucleic Acids Res 2002; 30:56–58.
66. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res 1999; 27:3821–3835.
67. Tao H, Bausch C, Richmond C, Blattner FR, Conway T. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J Bacteriol 1999; 181:6425–6440.
68. Helmann JD, Wu MF, Kobel PA, et al. Global transcriptional response of *Bacillus subtilis* to heat shock. J Bacteriol 2001; 183:7318–7328.
69. Merrell DS, Butler SM, Qadri F, et al. Host-induced epidemic spread of the cholera bacterium. Nature 2002; 417:642–645.
70. Oh MK, Liao JC. DNA microarray detection of metabolic responses to protein overproduction in *Escherichia coli*. Metab Eng 2000; 2:201–209.
71. Oh MK, Liao JC. Gene expression profiling by DNA microarrays and metabolic fluxes in *Escherichia coli*. Biotechnol Prog 2000; 16:278–286.
72. Loos A, Glanemann C, Willis LB, et al. Development and validation of *Corynebacterium* DNA microarrays. Appl Environ Microbiol 2001; 67:2310–2318.
73. Oh MK, Rohlin L, Kao KC, Liao JC. Global expression profiling of acetate-grown *Escherichia coli*. J Biol Chem 2002; 277:13,175–13,183.
74. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270:467–470.
75. Schut GJ, Zhou JZ, Adams MWW. DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: evidence for a new type of sulfur-reducing enzyme complex. J Bacteriol 2001; 183:7027–7036.
76. ter Linde JJ, Liang H, Davis RW, Steensma HY, van Dijken JP, Pronk JT. Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of *Saccharomyces cerevisiae*. J Bacteriol 1999; 181:7409–7413.
77. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997; 94:13,057–13,062.
78. Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. Cell 2000; 102:109–126.
79. Desprez T, Amselem J, Caboche M, Hofte H. Differential gene expression in *Arabidopsis* monitored using cDNA arrays. Plant J 1998; 14:643–652.
80. Kawasaki S, Borchert C, Deyholos M, et al. Gene expression profiles during the initial phase of salt stress in rice. Plant Cell 2001; 13:889–905.
81. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet 2001; 29:389–395.
82. Kaminski N, Allard JD, Pittet JF, et al. Global analysis of gene expression in pulmonary fibro-

sis reveals distinct programs regulating lung inflammation and fibrosis. Proc Natl Acad Sci USA 2000; 97:1778–1783.
83. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10,614–10,619.
84. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403:503–511.
85. Coller HA, Grandori C, Tamayo P, et al. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc Natl Acad Sci USA 2000; 97:3260–3265.
86. Chhabra SR, Shockley KR, Conners SB, Scott K, Wolfinger RD, Kelly RM. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J Biol Chem 2003; 278:7740–7752.
87. Demirjian DC, Shah PC, Moris-Varas F. Screening for novel enzymes. Top Curr Chem 1999; 200:1–29.
88. Dautin N, Karimova G, Ullmann A, Ladant D. Sensitive genetic screen for protease activity based on a cyclic AMP signaling cascade in *Escherichia coli*. J Bacteriol 2000; 182:7060–7066.
89. Yano T, Kagamiyama H. Directed evolution of ampicillin-resistant activity from a functionally unrelated DNA fragment: a laboratory model of molecular evolution. Proc Natl Acad Sci USA 2001; 98:903–907.
90. Koning SM, Albers SV, Konings WN, Driessen AJM. Sugar transport in (hyper)thermophilic archaea. Res Microbiol 2002; 153:61–67.
91. Bateman A, Birney E, Cerruti L, et al. The Pfam protein families database. Nucleic Acids Res 2002; 30:276–280.
92. Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. Nucleic Acids Res 1991; 19:6565–6572.
93. Huang JY, Brutlag DL. The EMOTIF database. Nucleic Acids Res 2001; 29:202–204.
94. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 2000; 28:231–234.
95. Attwood TK, Beck ME. PRINTS—a protein motif fingerprint database. Protein Eng 1994; 7: 841–848.
96. Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ. PRINTS—a database of protein motif fingerprints. Nucleic Acids Res 1994; 22:3590–3596.
97. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002; 30:281–283.
98. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a Web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 2000; 28:3442–3444.
99. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 278:631–637.
100. Tatusov RL, Natale DA, Garkavtsev IV, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001; 29:22–28.
101. Frishman D, Albermann K, Hani J, et al. Functional and structural genomics using PEDANT. Bioinformatics 2001; 17:44–57.
102. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 1998; 16:939–945.
103. McGuire AM, Hughes JD, Church GM. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res 2000; 10:744–757.

104. Fumoto M, Miyazaki S, Sugawara H. Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. Nucleic Acids Res 2002; 30:66–68.
105. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000; 28:304–305.
106. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 2002; 30:402–404.
107. Overbeek R, Larsen N, Pusch GD, et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 2000; 28:123–125.
108. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997; 8:581–599.
109. Hoffmann K, Stoffel W. TMbase—a database of membrane spanning proteins segments. Biol Chem Hoppe-Seyler 1993; 347:166.
110. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 2002; 30:47–49.
111. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673–4680.
112. Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999; 24:34–36.
113. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001; 305: 567–580.
114. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane α-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 1997; 10: 673–676.
115. Haft DH, Loftus BJ, Richardson DL, et al. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 2001; 29:41–43.
116. Baxevanis AD. The Molecular Biology Database Collection: 2002 update. Nucleic Acids Res 2002; 30:1–12.
117. Nelson KE, Paulsen IT, Heidelberg JF, Fraser CM. Status of genome projects for nonpathogenic bacteria and archaea. Nat Biotechnol 2000; 18:1049–1054.
118. Kawarabayasi Y, Hino Y, Horikawa H, et al. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 1999; 6:83–101, 145–152.
119. Direct submission: *Pyrococcus abyssi* Genomic Sequence. National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, 2001.
120. Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH. Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc Natl Acad Sci USA 2002; 99:984–989.
121. Deckert G, Warren PV, Gaasterland T, et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 1998; 392:353–358.
122. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 1999; 399:323–329.
123. Voorhorst WGB, Rik IL, Luesink EJ, Devos WM. Characterization of the celB gene coding for β-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus* and its expression and site-directed mutation in *Escherichia coli*. J Bacteriol 1995; 177:7105–7111.
124. Kengen SWM, Luesink EJ, Stams AJM, Zehnder AJB. Purification and characterization of an

extremely thermostable β-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Eur J Biochem 1993; 213:305–312.

125. Verhees CH. Direct submission: *Pyrococcus furiosus* Genomic Sequence. Laboratory of Microbiology, Wageningen University and Research Center, Wageningen, The Netherlands, 1999.
126. Bauer MW, Driskill LE, Callen W, Snead MA, Mathur EJ, Kelly RM. An endoglucanase, eglA, from the hyperthermophilic archaeon *Pyrococcus furiosus* hydrolyzes β-1,4 bonds in mixed-linkage (1->3),(1->4)-β-D-glucans and cellulose. J Bacteriol 1999; 181:284–290.
127. Bauer MW, Bylina EJ, Swanson RV, Kelly RM. Comparison of a β-glucosidase and a β-mannosidase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Purification, characterization, gene cloning, and sequence analysis. J Biol Chem 1996; 271:23,749–23,755.
128. Gueguen Y, Voorhorst WGB, van der Oost J, deVos WM. Molecular and biochemical characterization of an endo-β-1,3-glucanase of the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem 1997; 272:31,258–31,264.
129. Tanaka T, Fujiwara S, Nishikori S, Fukui T, Takagi M, Imanaka T. A unique chitinase with dual active sites and triple substrate binding sites from the hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. Appl Environ Microbiol 1999; 65:5338–5344.
130. Jorgensen S, Vorgias CE, Antranikian G. Cloning, sequencing, characterization, and expression of an extracellular α-amylase from the hyperthermophilic archaeon *Pyrococcus furiosus* in *Escherichia coli* and *Bacillus subtilis*. J Biol Chem 1997; 272:16,335–16,342.
131. Savchenko A, Vieille C, Kang S, Zeikus JG. *Pyrococcus furiosus* α-amylase is stabilized by calcium and zinc. Biochemistry 2002; 41:6193–6201.
132. Laderman K, Davis B, Krutzsch H, et al. The purification and characterization of an extremely thermostable α-amylase from the hyperthermophilic archaebacterium *Pyrococcus furiosus*. J Biol Chem 1993; 268:24,394–24,401.
133. Dong GQ, Vieille C, Zeikus JG. Cloning, sequencing, and expression of the gene encoding amylopullulanase from *Pyrococcus furiosus* and biochemical characterization of the recombinant enzyme. Appl Environ Microbiol 1997; 63:3577–3584.

26

基因组学在药物发现过程中的整合

Jacques Ravel

引言

在过去六十多年中，抗生素的发现成为医学领域非常突出的贡献，特别是在治疗细菌性感染疾病方面。这些化合物大多是从土壤微生物中分离得到的天然产物，尤其是链霉菌属的成员，目前使用的天然衍生抗生素中有三分之二来源于它们[1]。尽管迄今为止，已有为数颇丰的抗生素药物，但细菌性感染疾病仍是导致死亡的首要原因之一[2]。

20世纪末的传染病大爆发和流行，其中，既包括新型传染病，也包括那些曾经认为已经得到控制的旧传染病的再现[3]。有些微生物以前认为无毒害，而最近演化成强力的病原菌，例如，大肠杆菌及其食物携带大肠杆菌OH：157，这些病原菌的大爆发引起美国大量人口死亡[2,3]，这是对公共卫生严重威胁的一种趋势，并且由于抗生素抗性临床分离株的增长使问题更加复杂化。

使用（滥用）抗生素为病原微生物提供了强大的选择压，以致使这些病原微生物发展或获得了对许多抗生素的抗性，如今抗细菌药已成为第二大类处方药物[4]。抗生素使用的增加，加上社会和经济因素，使得临床医生的百宝药箱中最强力的抗生素功效开始下降。最近，在美国分离的对万古霉素产生抗性的肠道球菌和葡萄球菌[5]，对青霉素产生抗性的肺炎球菌[6]，以及对多种药物产生抗性的结核分枝杆菌（*Mycobacterium tuberculosis*）已经敲响了警钟[7]。

颇具讽刺意味的是，抗生素作为治疗药物的成功，已经阻碍了新型化学药物的发现和生产，反而在药物发现这个领域中最新的成功，是直接通过化学修饰改良现有抗生素（例如，第二代乃至第三代头孢菌素），或利用次级抑制剂，如克拉维酸（一种β内酰胺酶抑制剂，与β-内酰胺抗生素联用）。直到最近，美国食品与药物管理局批准了一种新唑烷类抗生素linezolid[8]，在这之前的30年里，没有新类型抗生素问世。令人不安的是，临床分离的金黄色葡萄球菌（*Staphylococcus aureus*），已经报道对linezolid有抗性，也对许多其他类抗生素，包括万古霉素产生了抗性[5]。

以上勾画的这张灰色图画，已经显著地突出了对新型抗生素的迫切需求。医药工业开始清醒地意识到新型疗法的必要性，因此，越来越多的兴趣聚焦在发展新技术手段，试图发现下一代抗生素药物。自1995年第一种微生物流感嗜血杆菌（*Haemophilus influenzae*）基因组测序完成后[9]，到目前为止已有170多种微生物基因组测序，并且至少有另外330个基因组的测序正在进行之中，见 http://wit.integratedgenomics.com/GOLD/。这种大规模的全球性测序主要集中在病原菌，几乎囊括了所有的基因组项目，并且产生了大量的原始数据用于计算机分析。而且，正在对同种生物的多个株系或者同

一个属内的多个种进行测序或已经完成测序，这样就为利用比较基因组学工具发现新型药物靶标打开了可能之门。后基因组时代的最主要挑战是发掘和解密这笔新的信息财富，以充分完成对新型抗生素急切需求的计划。

基因组学和生物信息学飞速的革命性发展，正准备在药物发现领域大显身手。如今，在寻找新型药物靶标过程中，基因组学已经成为药物发现流程中不可缺少的一部分，同时，它也可以应用于新型微生物天然产物的发现。本章将对这些新颖方面和技术作一综述。

药物发现和靶标设计

基因组学研究给新型抗生素及其他药物的发展带来新希望，如今可以用生物信息学工具分析基因组序列数据，并鉴定新药物靶标，而这些靶标很可能具有治疗潜能。工业上已经很普遍采用生物信息学技术分辨优先靶标，但这只是药物筛选漫长过程中的第一步，接下来的步骤包括，靶标确认、化验手段的发展、小分子库的筛选等，这些都是药物发现过程中最耗时费力的，从确定靶标到候选药物筛选之间可能要花好几年（图 1）。

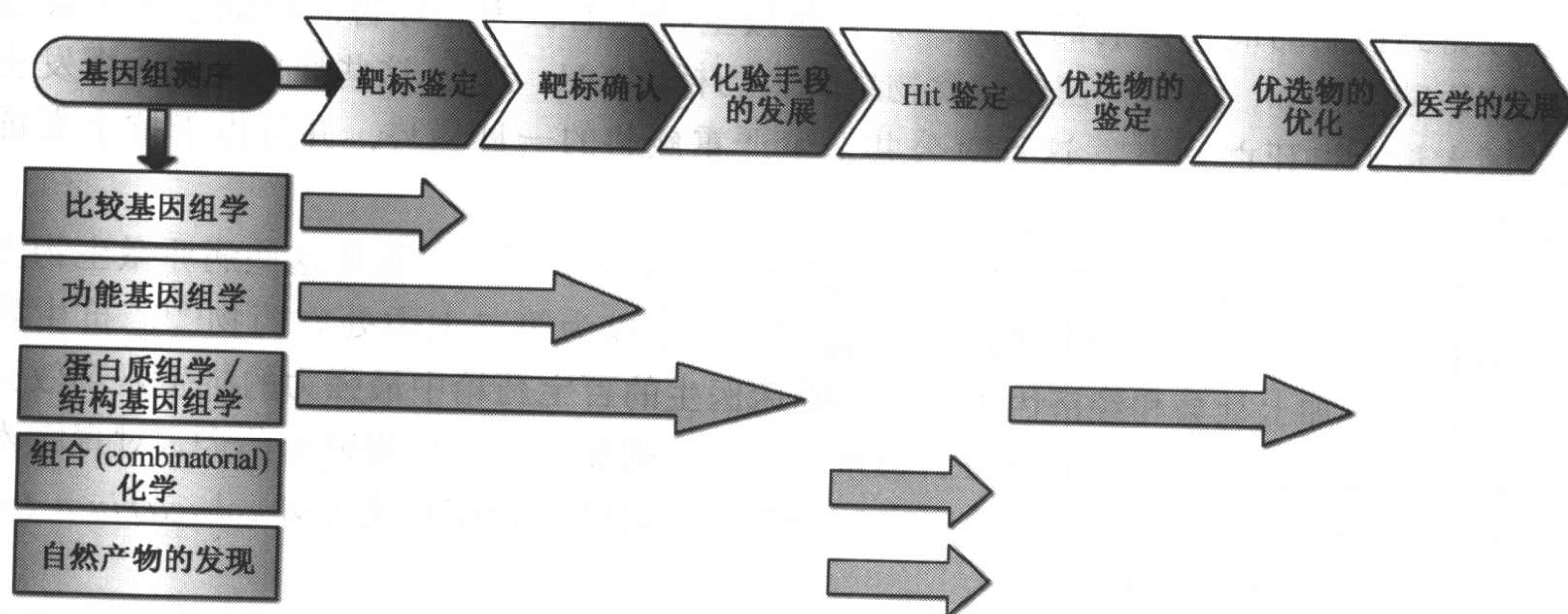


图 1　基因组学成为药物发现生产线中不可缺少的环节。从基因组测序到基因组分析以及结构基因组学，基因组学许多方面的不同阶段都要采用，从确定药物靶标到临床应用大约需要 10 年。

利用计算机分析鉴定靶标

如何定义一个“好”靶标基因

比较基因组学工具使生物信息学识别具有药物靶标潜力的可读框（ORF）成为可能，在文献中“好”抗生素靶标的标准已有了普遍的共识，那就是看它在必需性、选择特异性、广谱性、实用性以及功能性方面是不是优越。

必需性

药物靶标一定是实验微生物生长、复制和存活所必需的，因为抗生素活性的初级检

测是在这种条件下进行的，尽管如此，由于不同生长条件可能会导致特定基因功能的缺失，因此也要考虑一些其他标准。如果某一靶标是细菌在易感寄主中的存活所必需的，并需要利用该基因产物建立和维持感染[10]，那么，这个靶标具有必需性，这就说明毒性基因可以当作必需的靶标，但这仍受到质疑，因为很多基因的功能还不清楚[11]。

要证明并系同源物（paralog）是必需的，必须先把它们全部抑制，例如，革兰氏阴性菌螺旋酶 UvrD 和 Rep 的双突变株不能存活，而单突变株可以存活[12]，而生活在同样环境中的微生物直系同源物（ortholog）也不能作为靶标，因为替代物基因可以从一种非病原菌传递到一种病原分离物，从而迅速抵消药物的功能，这个例子说明，基因组技术只有与微生物生理知识紧密结合才能很好地发挥。

选择性

微生物靶标不应该具有与哺乳动物寄主保守性很好的同源物，以减少毒性问题。因此，最近人类和老鼠基因组测序的完成，使药物工业向前迈了一大步[13,14]。被选择靶标能代表该微生物的某一独特生物化学特性最为理想。

然而，针对独特的活性并不能保证对寄主没有毒性，举个例子，二磷酸尿嘧啶(UDP)-N-acetylglucosamine enolpyruvyl 转移酶（MurA），是负责细胞壁生物合成较早步骤的酶，与人类基因组没有任何同源性。然而，该酶的一个称为 fosfomycin 的抑制剂，却无法被寄主容忍[15]。另一方面，一种已经很完善的药物，thymethoprim 抑制二氢叶酸还原酶（合成 thymidylate 的关键酶，也是合成 DNA 的关键酶）[16]，尽管该酶在人类有一个同源物，人体对 thymethoprim 的耐受性很好，而又没有大的毒性[15]。

单单通过一级结构（氨基酸顺序）筛选潜在靶标，有许多局限性，因为，某些蛋白质尽管一级结构相似性很差，却有许多类似的蛋白质折叠，而这些折叠是抑制剂的结合位点。因此，在评估药物选择性时应同时考虑蛋白质的二级和三级结构。

广谱性

根据在微生物中的分布和潜在的抑制剂谱性，药物靶标可以分为两类：第一，存在大量病原菌中的靶标在广谱药物发现中有用。纵观临床在快速鉴定感染病原体面临的挑战，用广谱制剂进行治疗更恰当，成本效率比更合算，且便于临床应用[17]。第二，靶标只对某一种或一小类病原菌有特异性，以发展窄谱药类。窄谱药物在慢性感染或长期治疗中极其有用，它可以使有害效应降到最小，特别是对正常肠道微生物菌系，而且，窄谱抗生素可以减少抗药性向其他病原菌传递[18,19]。

实用性

利用生物信息学工具一旦选择了某种蛋白质，就要花费相当长的时间去测定它是否是一个可行的靶标。必须发展新检测手段，对蛋白质进行高通量筛选，生物化学方法常常依赖于可溶性蛋白，这样某些种类的蛋白难以用作靶标。例如，在体外过量表达的膜蛋白属难溶性蛋白，这就为检测方法的发展提出了特殊的挑战。

暴露在表面的膜蛋白，如果必需，代表了几乎完美的靶标。药物攻克这样的靶标不需要穿透细胞，因此简化了化学设计。这样，药物分子质量可以大一点，而且容易进行

优化。然而，正如前面提到的，测定药物效价有困难，因为检测方法的发展由于体外表达和可溶性问题而受阻。很多抗生素的靶标定位于膜蛋白，例如，细胞壁生物合成机器是通过全细胞测定方法发现的，它们的特异性靶标随后才阐明。在病原菌中有很多膜运输系统，因为这些微生物依靠它们的寄主提供必需的营养，这些也是颇具吸引力的靶标。然而，定位于这些运输系统的药物靶标可能是对人类有毒[11]。一个例外是衣原体和立克次氏体的 ATP/ADP 转位酶，它们与植物叶绿体更接近，而不是人类的线粒体[20]。

功能

关于靶标功能的信息对生产可行的化学药物十分关键，例如，功能信息可以帮助设计化学药物文库去筛选对靶标的抑制剂。结构基因组学的出现成为这个过程的关键，而且，生物化学与生理学信息结合结构方面的数据可以极大地提高成功率。

基因组比较与靶标优选

如上所述，确定靶标所用的一系列参数会影响基因组序列数据分析的过程，这些参数可能适用于整个基因组数据以及相应的靶标优选，促进这一过程的生物信息学工具在私人或公共机构都有所发展，这些工具往往通过利用基因组序列数据库查询所需信息，这种查询通常基于蛋白质序列的相似性，用 BLASTP（Basic Local Aligment Search Tool P）运算法进行[21]。

其他类型的比较分析则用功能模体（motif）的保守性以及隐式马可夫模型[22,23]、PSI-BLAST（Position-specific iterated BLAST）、PHI-BLAST（Pattern-Hit initiated BLAST, 24）或 Clusters of orthologous Group（COG）[25]。有些是采用酶学委员会分类法，把那些序列相似性差，但功能相近的蛋白质根据 EC 序号加以鉴定[26]。用于比较基因组学的 HOBACGEN（Homologous Bacterial Genes）数据库，可以使用户筛选基因家族，并查出数据库中分类同源性，这对鉴定直系同源和并系同源基因很有帮助[27]。

基因组研究所的微生物资源大全（CMR），为完整的微生物基因组序列提供了详尽的数据库（http: www.tigr.org/tigr-scripts/CMR2/CMRhomepage.spl）。Read 等[18]利用 CMR 的上呼吸道病原菌，如肺炎链球菌（*Streptococcus pneumoniae*）、流感嗜血杆菌和脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）的基因组，筛选定点治疗的可能基因。对这些生物的比较也可以进一步了解呼吸道中微生物生存的基本策略。利用蛋白质序列的同源性（BLASTP）（域值为氨基酸相似性在 25%），他们鉴定了与这三种病原菌有共同性的 32 种蛋白，而这些蛋白质在非病原菌株大肠杆菌 K12 和枯草芽孢杆菌中不存在，这种分析揭示了一些功能未知的基因，以及一些已知在其他病原菌中保守的毒性因子，同时也为公众提供了利用网上数据库进行生物信息学分析的范例。

与之类似，Bruccoleri 等利用和谐分析法（condordance analysis）与 COGs 数据库为广谱抗生素鉴定了 89 个靶标[28]，这些基因同时在大肠杆菌、枯草芽孢杆菌、流感嗜血菌、幽门螺杆菌（*Helicobacter pylori*）和结核分枝杆菌中存在。这 89 种蛋白包括一些有用的抗菌靶标，例如，DNA 促旋酶（GyrA）、quinolone 类抗生素的靶标[29]，还有 fosfomycin 的靶标——murA[30]。

Chalker 等[31]用另一种比较基因组学方法，鉴定幽门螺杆菌两个株系——26695 和 J99 之间保守的可读框（ORF）和另外 13 种高度分歧的真细菌。利用一系列筛选参数，例如 BLASTP 得分、疏水特性以及推定的功能分类，他们鉴定了 73 个可读框，并认定其符合“好”靶标的定义。其中，跨越几个功能类别的 45 个可读框，被选作必需性测试，利用快速无载体等位互换突变技术（vector-free allelic replacement mutagenesis）的体外测试，发现 33 个可读框具有必需性。有趣的是，发现了 12 个必需基因，在其他细菌中的直系同源基因则是非必需的。这些高度差异的非必须基因产物，可能是新颖而有高度特异性抗幽门螺杆菌药物的潜在靶标，用直接与抗菌药物发现有关的简单生物标准对可读框进行逐步优选，可能为筛选潜在靶标基因和鉴定它们对正常微生物细胞功能的关键性提供有效的途径，这种分析可以产生相当小的一套基因，去验证其必需性，提高下游处理过程的成本效率。

由于对微生物生理的了解还有很大差距，生物信息学分析会产生一大套没有功能分配的可读框，这些基因中有一些好靶标。大多数药物发展项目需要了解或预测蛋白质功能，作为它成为潜在靶标的前提。因此，一大套潜在靶标在进一步研究之前可能被淘汰，而在测序的基因组中，一般有三分之一的可读框没有功能定位。

然而，基因组学技术已经发展到可以从未知蛋白中提取功能信息，基因芯片转录图谱可为未知基因的定时表达提供信息，通过变化环境生长条件（如培养基、在寄主中的生长或暴露在抗生素中），可以收集很多有价值的信息，而这些信息可以直接导致功能定位，因为表达模式共享的基因，功能也共享。

然而，并不是所有的细胞过程都在基因表达水平受调控，蛋白质组学可为基因功能的评估提供互补工具，尤其是非常灵敏多样的质谱技术，可以在基因组范围尺度内分析蛋白，这些技术可以用来评估蛋白质在不同生长条件下的蛋白质水平[32~34]。除了可以用于鉴定未知基因的功能外，这些转录组和蛋白质组技术还可以用来鉴定和评估新颖靶标[15]。

检验必需性

为了考虑进一步研究，通过生物信息学分析得到潜在靶标，需要进行生物筛选检验必需性。正如靶标鉴定一节中提到的，蛋白质必须是微生物生存所必需，才能作为有效靶标，基因的必需性可以通过体外实验或动物模型的体内感染实验进行鉴定。

长期以来，体外必需性一直用作检验抗微生物靶标，用鼠伤寒沙门氏菌（*Salmoella typhimurium*）温度敏感突变株已经发现，一系列基因在实验室富集培养基中的生长是必需的[35]，这些方法是基于随机突变株库的产生，接着进行突变位点的遗传图谱定位，这是事倍功半的。

当潜在药物靶标通过生物信息学分析鉴定后，序列定向法为必需基因的鉴定提供颇具吸引力的另一途径。

质粒插入突变利用含筛选标记的自杀质粒，将靶标基因的一部分克隆到质粒上，该质粒转化到一个非允许的寄主中。基因失活是通过质粒上的基因与染色体位点之间的同源重组实现，鉴定金黄色葡萄球菌生长必需基因就是利用这种方法[36]，在肺炎链球菌中已经获得高通量的基因失活，检验了 347 个可读框，发现其中 113 个基因是必需

的[37]。然而，这些插入失活技术也受到一些干扰，例如，质粒插入操纵子引起的极性效应，被剪切的蛋白质仍保持活性，由质粒切除引起的回复突变。

另一种方法，称为等位互换突变，则可以产生稳定的完全基因缺失，而不会产生有害的极性效应。等位互换构体，包括一个可筛选的标记，其侧翼与靶标位点上下游同源的序列，该构体可以直接导入野生型细胞中，也可通过自杀质粒在两侧翼同源区产生双交叉同源重组而进入细胞，靶位点基因便可被筛选的标记物替代。另外，可以用多个重叠 PCR 大量构建突变体，选择 PCR 引物使侧翼基因和潜在启动子在缺失突变株中保持不变，从而减少极性效应。如果三个或更多独立的转化不能产生某基因的突变株，该基因即可定为必需基因。

等位互换突变被 Chalker 等用于鉴定幽门螺杆菌生长必需的基因[31]，类似的研究还有 Wilding 等鉴定了肺炎链球菌中生存必需的 3-hydroxy-3-methylglutaryl-coenzyme A synthase，该合成酶是甲羟戊酸途径中控制革兰氏阳性球菌二磷酸异戊烯的生物合成，并与红霉素（erythomycin）一个抗性片段的等位互换，致使微生物对甲羟戊酸的营养缺陷，并使老鼠呼吸道感染模型中的毒力严重消弱[38]。因此，利用基因组优选以及体内和体外靶标确认的方法，我已证明，参与二磷酸异戊烯生物合成酶是革兰氏阳性球菌中一个好药物靶标。

结构基因组学

结构基因组学是“omics”领域中的新成员，正如在前面有关章节所提到，基因组时代最大的挑战是挖掘基因组测序提供的宝贵信息财富，结构基因组的最终目标是为每一种蛋白质提供三维结构信息[39]。已完成的众多微生物基因组，由于它们的结构相对简单，为实施这一目标提供了很好的模板，结构信息可为未知蛋白的生物功能定位，并在其他方面为药物发现过程提供帮助，因此，它对科学家的指南作用是无价的。Gilliland 等[40]利用结构基因组学，开展了流感嗜血菌未知基因产物的功能定位，并发现了一系列新药物靶标。

截至 2002 年 10 月 1 日，在公共蛋白质数据库（PDB）中，已经储藏了 18 000 多个三维结构信息[41]，然而，只有 5000 个这样的结构代表了野生型蛋白质，其余的全都是突变体、复制品或酶-配体复合物。

解析一种特别蛋白的结构，历来是缓慢而耗时的，难以成为高通量自动化系统。然而，最新发展的基因组测序、蛋白质表达[42]、selenomethionine 标记技术、多波长不规则衍射定相和同步加速器结晶，已极大地促进了蛋白质结构测定的过程[43]。尽管 X 射线结晶学在精细分辨结构方面具有明显优越性，而核磁共振（NMR）技术可对蛋白质的天然状态进行分析，没有像晶体形成那样的限速步骤[44,45]。

总之，蛋白质结构在分辨率和质量方面的提高非常迅速，尽管还面临很多挑战，具有不同背景的科学家们正共同努力，使高通量结构测定能成为现实。计算机技术和生物信息学发展，也为三维结构分析获得的数据处理提供必要的工具[46,47]。

结构基因组学的技术难题

三维结构测定需要大量可溶性蛋白质，蛋白质的量是结构基因组项目的关键，大量

的精力集中在蛋白质表达和纯化过程的自动化，在未来几年中，一个小研究组就有望每年表达和纯化数以万计的蛋白质。大量功能性蛋白质生产技术的发展以及随后的结构分析，也会造福于其他领域，例如，蛋白质芯片技术，以及为药物发现发展的筛选测试。蛋白质在大肠杆菌中表达和纯化技术的优化（如质粒载体的选择、纯化标记、培养基等），也受益于高通量技术和处理过程的小型化[42]。

尽管已做出种种努力，据估计，仍有三分之一到一半的已知原核生物蛋白质，不能表达成可溶形式[48]。蛋白质表达和纯化问题给结构基因组和靶标筛选提出的一个主要挑战，因为这两个过程都需要大量可溶性蛋白。一种解决办法是在平行表达和纯化实验中，筛选给定蛋白质的最可溶直系同源物[49]。在大肠杆菌表达系统中，我们发现 80% 的研究蛋白，至少有一个，经常有多个可溶性变体[50]，这就增加了 X 射线衍射研究中获得晶体的概率，而且可以单独构建并表达蛋白质的各个结构域，这样就潜在地提高了可溶性以及晶体衍射的质量。实验分析（如限制性蛋白质水解与质谱联用的方法）比序列比较更好地提示了结构域的界限[51]，生物信息学工具正在发展并有可能准确推断结构信息[52, 53]。

正如在实用性章节中提到的，膜镶嵌蛋白（如运输系统及细胞囊的生物合成组分）提供了颇有希望的抗菌药物靶标，然而，因为它们比较差的可溶性，常常被排除在药物靶标库之外，膜蛋白是 X 射线晶体学和 NMR 分析中最大的挑战之一。新技术，如脂立方相介导的结晶（lipid cubic phase-mediated crystallization），是这方面很有希望的技术[54,55]，如果该技术成熟，蛋白质数据库中膜蛋白结构域的数量将会迅速增加[41]。

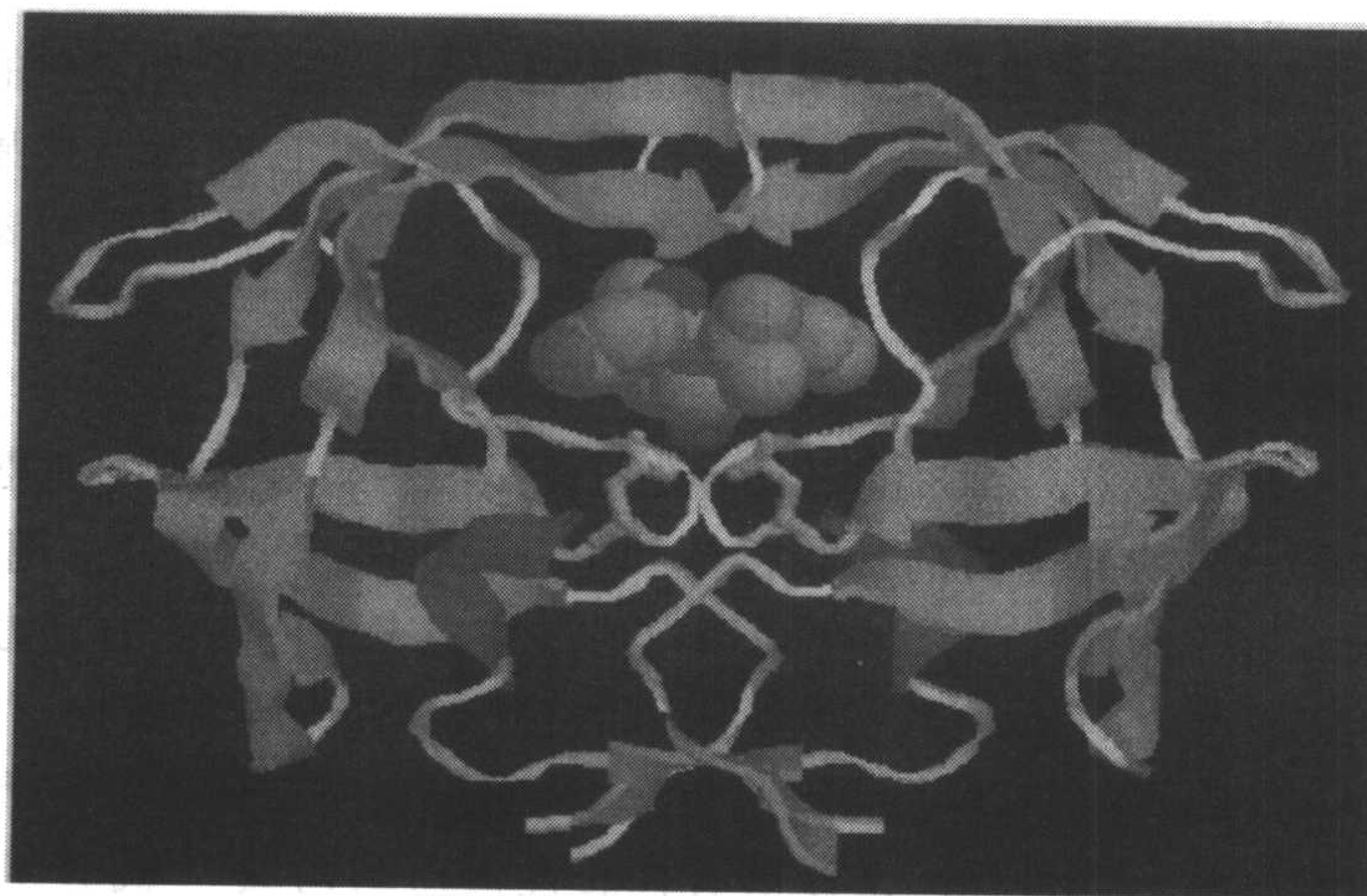


图 2　基于蛋白质结构的药物设计。图中球状分子代表 Amprenavir，它是一种人类免疫缺乏病毒蛋白酶（带状）的抑制剂，该分子结合在蛋白酶的活性位点。

结构基因组学在药物发现过程中的整合

现在，医药工业已记录了利用蛋白质晶体结构设计药物，例如，利用合理的结构技术设计酶抑制剂，然而，这种应用已经受到限制，并在获取高质量蛋白质结构信息方面遇到难题。除了靶标发现外，结构基因组学将对药物发现过程产生至少两方面的影响：

(1) 优选物 (lead) 的鉴定，即针对一种特别靶标而构建和筛选活性小分子化合物；(2) 优选物的优化，即对筛选出的优选物分子进行化学精细加工，以提高它的选择性和药效 (图 1)。在这两个方面，生物靶标或相关蛋白质的三维结构，将允许科学家以更合理和有效的方法产生优选物分子[39]。基于结构的药物设计已经成功生产了市场化药物，这都源于结构和功能分析，这些药物包括，neuramidase 抑制剂 zanamivir (Relenza)、人类免疫缺乏症病毒蛋白酶抑制剂 amprenavir (Agenerase) 和 nefelvir (Viracept)[43]，以及 Bcr-Ab1 酪氨酸激酶抑制剂 imatinib (Gleevec)[56] (图 2)。

靶标的功能信息对治疗设计极为重要，对测试、筛选和确认也十分关键。首先鉴定功能是后基因组时代药物成功发现的关键[57]，因为蛋白质功能在很大程度上是三维形态的结果而不是氨基酸顺序，结构基因组技术的发展将在药物发现中起关键作用。通常根据 DNA 或氨基酸序列的相似性推断功能。这可能导致对基因或可读框的错误命名和对功能的错误解释，众所周知，氨基酸序列相似的蛋白质可以有不同的三维结构因而有不同的功能 [58]，增加蛋白质三维结构库的容量将最有可能为大量已知蛋白提供结构模板。

当从序列分析中看不出明显的蛋白质折叠时，就要从整个体系分析中 (如同源蛋白质的折叠和聚合水平)[49]收集生物化学和功能方面的信息。大肠杆菌 FabH ketoacyl 载体蛋白合成酶活性位点的鉴定，就是通过另一种凝缩酶类似折叠的比较而获得[59]。一些局部特性，如极性残基决定簇或疏水片段、净电荷以及束缚态小分子配体和金属离子，都可用于鉴定功能位点、表面特性，如形状以及静电学性质也可用于推断功能特性[49]。

原则上，这些方法的组合在基因组测序项目中已经用于阐明未知蛋白的功能 [60~63]，结构基因组学所阐述的 LuxS 功能，这是一种保守蛋白，它参与群体感应 (quorum-sensing) 中双组分信号传导，以及一些革兰氏阴性和阳性病原菌中毒性基因的表达 [64]，在生物合成必需自身诱导分子 AI-2 时，LuxS 则以同源二聚体形式将 S-ribosylhomocysteine 的核糖环剪切。

这种方法并不常常奏效，因为蛋白质折叠可能是一种新颖的或特别类型的折叠并具有多项功能，如磷酸丙糖异构酶的桶状折叠和 Rossman 折叠。然而，随着愈来愈多蛋白质结构在数据库中的储存，建立预测模型以通过结构来定位功能的能力将会提高。根据网络的一些工具，如 Gen3D [65] 可以利用 CATH 结构分类[66]蛋白 (class, architecture, topology, homologous)，以在整个基因组水平对蛋白质结构进行解析。

结构基因组学的另一个目标，是提供有代表性蛋白结构域[49]高分辨率的模板，这些模板可以用于推测蛋白质折叠、结构及最终功能。序列-结构同源性模型正在发展中，一些软件，如 FUGUE 在蛋白质折叠预测方面正取得很大成功 [67,68]，这些预测模型可以运用到药物设计的靶标发现中。鉴定某病原生物特异性蛋白质的折叠非常有价值，即使蛋白质之间有很强的氨基酸序列相似性，它们的三维结构也可能不同。因此，如果利用简单的氨基酸序列相似性比较，那么一些独特的折叠和潜在药物结合位点就可能被错过，同样，有些蛋白质可能只有有限的氨基酸相似性，但它们仍能采用相似的三维折叠，从而具有相似的小分子结合位点。

因此，结构基因组学技术的进一步发展，可望为药物发现和设计提供新的机会。目

前，NMR 和晶体技术已经常规地应用到探测配体结合相互作用，为筛选结合功能未知蛋白的小分子提供了另外的途径。通过筛选生物化合物库（如辅助因子和抑制剂），这一类型的方法可以获得未知蛋白质靶标的功能信息。

结构基因组学，尽管还处在婴儿期，但它正快速发展并成为药物发现过程中不可分割的一部分。整个蛋白质组的结构信息将帮助解决药物代谢和毒性问题，一些诸如配体库（ligand depot）数据库的计划，将会涵盖 250 000 多个小分子的信息，与其他蛋白质结构数据库的结合非常重要，因为越来越多的蛋白质结构将得到解析 [69]。

新天然产物的发现与基因组学

基因组学不仅可以运用到药物发现中，它还可以帮助寻找新天然产物，如药物发展过程中的优选化合物。基因组学在药物发现中的运用正在迅速改变医药工业，在过去 60 年里，医药工业已经发掘出生物合成的能力，利用微生物生产抗菌药物。

历史上，抗生素定义为低分子质量的有机天然产物，次生代谢产物，由微生物生产极低浓度抗另一微生物的物质 [70]。化学的发展已经在某种程度上改变了这个定义，但微生物仍然是医药抗生素的主要来源，许多医药公司正逐步停止寻找微生物次生代谢产物，而倾向于用组合化学（combinatorial chemistry）产生的化学分子库，尽管事实上该技术还没有把一种新药推向市场。

次生代谢产物已发展到可以应对自然区域环境的需求和挑战，大自然正在不断地产生自己的组合化学 [70]。应该记住，微生物有庞大的多样性，其中只有很小比例被检测过并生产次生代谢产物。大自然的化学生物多样性还很少被挖掘，据估计，只有 1%～10%的微生物根据其生存环境可以培养，很多计划已经开始挖掘这笔财富，下面列举了利用微生物基因组进展发起的几个计划，这些新方法并不局限于抗菌药物的发现，而且还包括发现抗癌、抗炎症、免疫抑制剂和抗寄生虫的次生代谢产物，以及除草剂和杀虫剂。

基因组扫描

一种高通量方法叫基因组扫描，在不依赖基因表达的条件下，搜寻次生代谢产物基因座[71]，该方法利用的依据，就是需要次生代谢产物生物合成的基因经常在细菌基因组上聚成簇（图 3），用这种方法，在任一微生物的所有天然产物基因簇都能克隆和测序，并不需要整个基因组的完全测序就可以进行分析。

到目前为止，采用基因组扫描法已发现了 400 多种天然产物基因簇，它们编码了结构不同的生物活性分子，包括抗生素 everninomicin（一种寡糖）、ramoplanin（一种糖基化的脂肽）、rosaramicin（一种聚酮）以及抗肿瘤抗生素 calicheamicin（一种 enediyne），还有 anthromycin（一种 benzodiazepinone）等 [72～76]。对放线菌中至少 10 属的 50 多个菌株的分析发现，在一种生物体内找到一打或更多天然产物基因簇是十分典型的，甚至对那些通过发酵筛选认为只有一种天然产物的生物也是如此。

因此，尽管放线菌生物合成的天才多年来一直受到赞赏，基因组扫描表明，它们生产天然产物的能力已被大大低估了，这些发现很好地展示了新天然药物发现的源泉，以及基因组学在新生物活性分子发现中起越来越重要的作用。

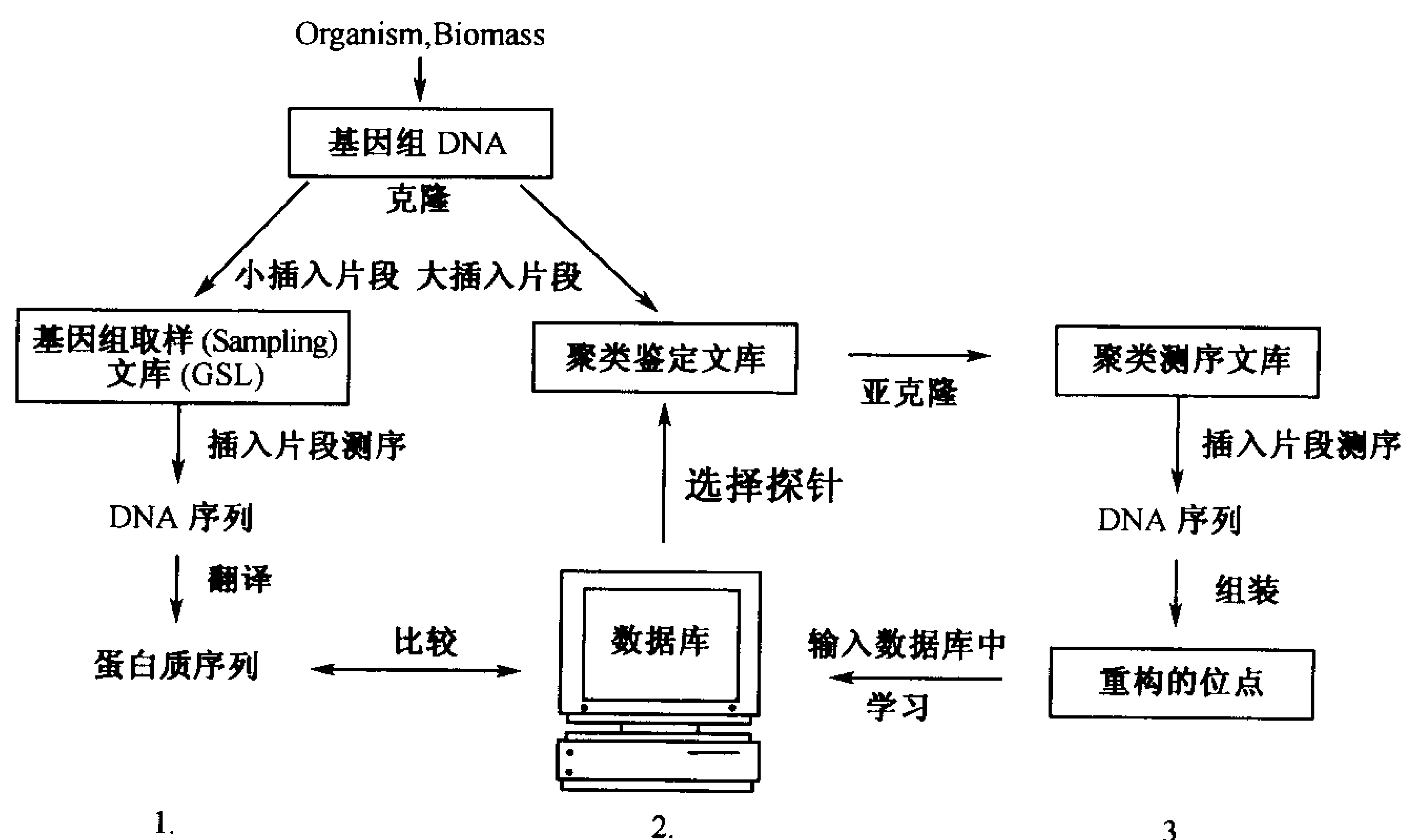


图 3　基因组扫描法在微生物基因组中发现天然产物基因簇的流程图[71]。基因组扫描为发现天然产物基因簇提供了有效途径，而不需要全基因组序列测定。该方法的依据，是天然产物的合成基因在细菌基因组上聚集成簇。

①高分子质量的基因组 DNA 被随机片段化，小片段用于制备以质粒为载体的基因组样本库（GSL)，大片段用于制备以黏粒或 BAC 为载体的基因簇鉴定库(GIL)。利用定位于质粒载体上的通用引物，可以从 GSL 克隆中获得有限数量的基因序列，用这些序列与数据库中天然产物生物合成基因比较，以鉴定含天然产物生物合成基因的克隆。

②用生物信息学分析找到的基因作探针，寻找那些含有相应基因的 GIL 克隆，这些基因及它们邻近的基因可能组成生物合成基因座。

③筛选出的 CIL 被随机地片段化，为准备第二个质粒库提供下一步测序的模板。测序和组装所筛选的 CIL 克隆结果，会产生一个完整的天然产物基因簇，然后命名和输入数据库。该方法的效率随数据库容量的增大而提高，因为，如果数据库中含有大量基因簇，那么，未知基因簇通过一定数量的 GSL 克隆分析而识别的概率就提高了。DecipherTM 数据库（Ecopia BioSciences，Inc.，Saint-Lautent，Quebec，Canada）最初从公共数据库（如 Genbank）中，收集能代表不同范围天然产物类型的基因簇，随后用基因组扫描发现的新基因簇丰富该库的内容。

探索环境泛基因组，发现新颖天然产物

培养微生物已经产生了一些有生物活性和医学用途的化合物，然而，重新发现已经鉴定的分子正在限制对可培养微生物的利用[77]，正如前面提到的，这些生物代表了环境中微生物多样性的很小比例，如果这个多样性是未能培养微生物的一个化学能力指标，许多新化合物还有待发现（见 23 章)。

为开发这些未被利用的代谢产物，环境中的 DNA 可以克隆到 BAC 载体上，后者可以稳定地维持大片段 DNA（＞100kb)，因而，可以利用环境群落中的泛基因组(metagenome)[78]。泛基因组代表了一个给定环境样品中可培养和未能培养微生物的遗传和功能多样性，这些泛基因组的 BAC 文库用在异源寄主菌（如大肠杆菌、枯草芽孢

杆菌和链霉菌）中筛选新的表型，在不同寄主中筛选 BAC 文库可望提高产生新活性的概率。

这种独立培养的基因组方法，已经成功地用于抗菌药物的发现，一般情况下，次生代谢产物生物合成基因与必需的自我抗性基因是聚集成簇。BAC 的利用增加了寄主表达整个途径的概率，一旦检测到新活性，遗传材料可立即用于分析和操作。

通过筛选环境中 DNA 库的克隆，已经分离到长链 N-acyltyrosine 衍生天然产物的两个新家族，这些克隆具有合成生物活性化合物的能力[79]。对这些克隆的序列分析以及通过转座子敲除各个基因的结果表明，由 13 个 ORF 组成的基因簇负责这一活性，用变铅青链霉菌（*Streptomyces lividans*）作为异源寄主，一个新家族的化合物——terragine 从土壤泛基因组库中发现[80]。

这些报道表明，不依赖于培养而直接通向环境化学多样性的方法，具有揭示新生物合成活性的潜力。作为这一原则的例证，violacein 是一个已知的广谱抗生素，就是由携带有土壤泛基因组库的大肠杆菌产生[81]，它的生物合成基因簇已经被研究清楚了。最新利用泛基因组发现的 tubomycin，表明这个方法为新的活性的发现创造了一个环境 DNA 和寄主的遗传材料之间的组合途径，因而强化了寄主的生物合成的能力。

Courtois 等[82]用 PCR 方法成功地筛选到一个小土壤泛基因组库（有 5000 个克隆），以获得聚酮生物合成基因，其中某些基因在变铅青链霉菌中异源表达时可以产生新型分子。这些数据进一步说明，发掘前所未知或未能培养微生物去发现新型天然产物具有潜在价值，而最重要的是，为把这种技术开发成一种现实而有效的药物发现工具提供了战略思路。

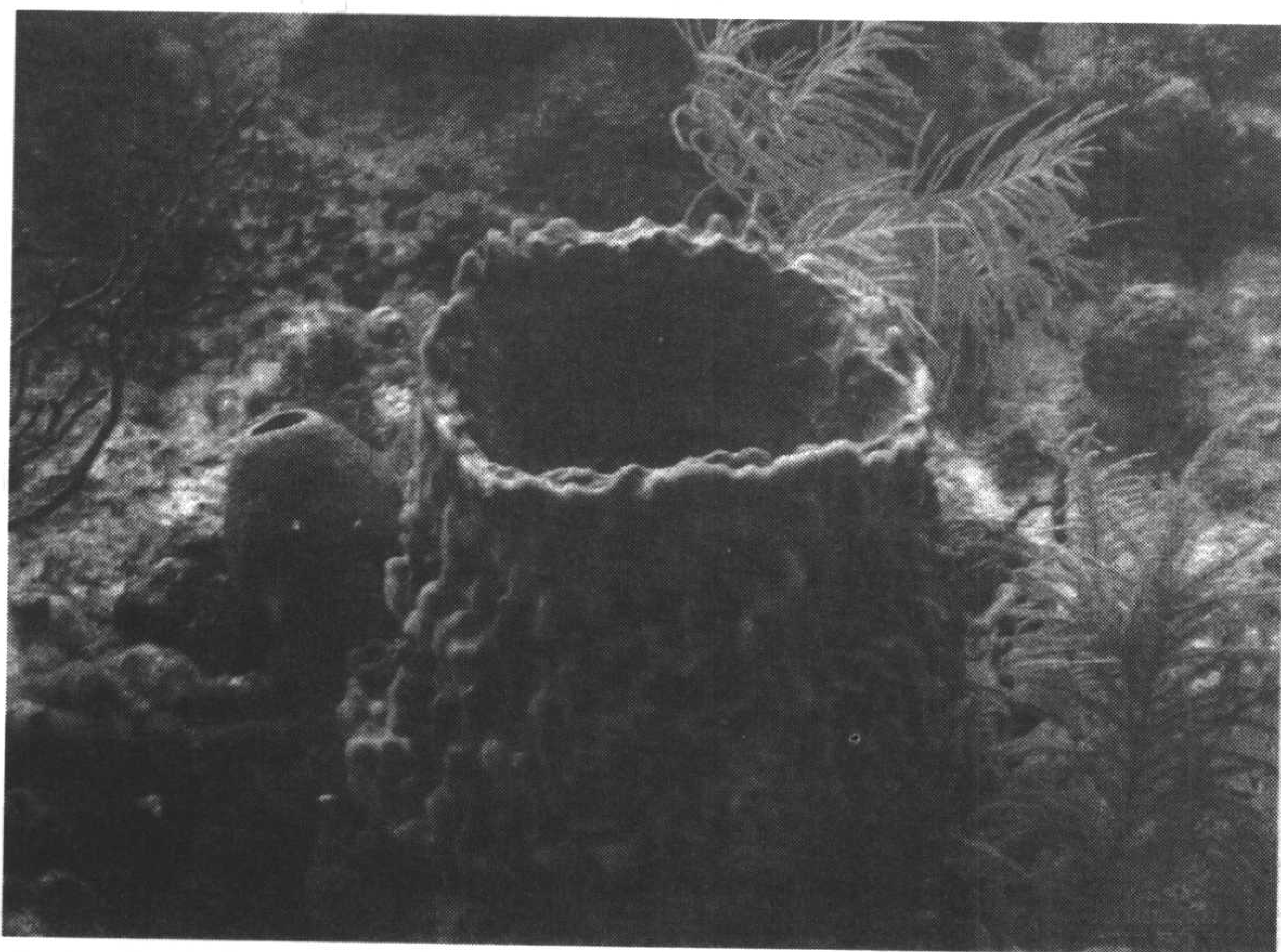


图 4　下一个重磅炸弹式的新药物会不会在海洋环境中发现？已从美国佛罗里达州 Key Largo 海洋中的海绵类生物，如 *Xestospongia muta* 中，提取了大量生物活性物质，基因组学可为这种资源的利用提供帮助（该图由 Jayme Lohr 提供）[80]。

泛基因组鉴定新天然产物到目前为止只用于土壤环境，以前一直是把可培养微生物

作为发现新化合物的丰富源泉[77]。然而，人们越来越重视从海洋微生物中筛选天然产物，这些产物中的一部分最初从海生无脊椎动物中发现[83]，实际上很可能是由与之共生的微生物产生的，海洋环境为这些泛基因组方法提供了发现新型天然产物的肥沃土壤，而且，从海洋环境生物中得到的这些化合物可提供而在陆生资源中没有的新型结构[83]。

在很多情况下，培养海洋无脊椎动物的共生微生物不大可能（图 4)[84]，泛基因组为获得这些次生代谢产物提供了现实途径，虽然目前泛基因组的应用只局限于生物工程领域，但它最终能为医药工业提供新化合物的来源。

用基因组改组技术改造天然产物

DNA 改组技术使基因和亚基因组 DNA 片段快速进化[85,86]，该技术是原生质体融合和 DNA 在多亲之间相结合并改组，以改善细菌的表型[87]。传统株系的改善由随机突变和筛选的连续轮回来指导微生物进化，以筛选到改良的表型，通过这些，在每一轮中筛出最好的突变株，以图进一步完善。基因组随机拼接利用第一轮诱变产生的遗传多样性，产生的新性状优良的突变株种群，并在下一轮中继续改组。

该技术已用于快速改良弗氏链霉菌（*Streptomyces fradiae*）产生的泰乐菌素，并表明经两轮基因组改组（一年）获得泰乐菌素产量，相当于传统方法 20 年所取得的成绩[87]。类似的其他表型也可以提高，例如，工业化乳酸杆菌（*Lactobacillus*）一个菌株通过基因组改组表明酸耐受性提高[88]，对 pH 值耐受性是由至少 18 个基因座 60 多个基因调控的，因此，这种改组方法的实用性和意义就显而易见了。

可以看出，这项技术可以用来改造新型次生代谢产物，如改良抗生素，而且，可以通过组合整个基因组的生物合成途径获得新的结构，这些新基因组技术正在成为药物发现的新工具。

结论

基因组时代正在提供大量潜在靶标，这些靶标需要各种药物化学多样性，组合化学技术的快速发展，将毫无疑问提供了某些多样性，然而微生物天然产物仍有光明前途，因为多种新型基因组技术正在发展，以开发自然界中无限的化学多样性。天然产物的发现还要进一步深入，因为许多环境还没有被尝试，很多生物还有待于培养，与病原微生物的战斗需要从许多方面进行。

基因组学对药物发现有广泛的影响，基因组学和生物信息学技术使感染的遗传基础和病原性进一步阐明，靶标的鉴定可以用于筛选新药。这些药物发现的新方法，正在扩展抗感染病的处理选项和抵御策略。在过去十年中，化学家、生物化学家、生物学家、微生物学家、结构学专家、计算机专家、医生、药理学家和基因组学家共同努力，解决药物研究面临日趋复杂的问题，这样一种多学科交叉的方法，将使得在抵抗病原微生物的战斗中取得前所未有的进展。

基因组学成为新药发现中具有创造性的发展动力，医药工业正有效地把这些新技术整合到药物发现中，尽管这个领域还很年轻，最新进展表明，基因组正在为提高人类健

康做出巨大贡献。

致谢

感谢 Ecopia Bioscience, Inc. 的 Chris Farnet 提供图 3 以及对手稿的建设性意见；也感谢 Brian Dougherty 对手稿的严格审阅和建议。

（吴天福 译）

参考文献

1. Bentley SD, Chater KF, Cerdeno-Tarraga AM, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 2002; 417:141–147.
2. Cassell GH, Mekalanos J. Development of antimicrobial agents in the era of new and reemerging infectious diseases and increasing antibiotic resistance. JAMA 2001; 285:601–605.
3. Cohen ML. Changing patterns of infectious disease. Nature 2000; 406:762–767.
4. McCaig LF, Hughes JM. Trends in antimicrobial drug prescribing among office-based physicians in the United States. JAMA 1995; 273:214–219.
5. *Staphylococcus aureus* resistant to vancomycin—United States, 2002. MMWR Morb Mortal Wkly Rep 2002; 51:565–567.
6. Geographic variation in penicillin resistance in *Streptococcus pneumoniae*—selected sites, United States, 1997. MMWR Morb Mortal Wkly Rep 1999; 48:656–661.
7. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. Emerg Infect Dis 2002; 8:843–849.
8. Shinabarger DL, Marotti KR, Murray RW, et al. Mechanism of action of oxazolidinones: effects of linezolid and eperezolid on translation reactions. Antimicrob Agents Chemoth 1997; 41: 2132–2136.
9. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
10. Buysse JM. The role of genomics in antibacterial target discovery. Curr Med Chem 2001; 8: 1713–1726.
11. Galperin MY, Koonin EV. Searching for drug targets in microbial genomes. Curr Opin Biotechnol 1999; 10:571–578.
12. Petit MA, Ehrlich D. Essential bacterial helicases that counteract the toxicity of recombination proteins. EMBO J 2002; 21:3137–3147.
13. Mural RJ, Adams MD, Myers EW, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 2002; 296:1661–1671.
14. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291:1304–1351.
15. Haney SA, Alksne LE, Dunman PM, Murphy E, Projan SJ. Genomics in anti-infective drug discovery getting to endgame. Curr Pharm Des 2002; 8:1099–1118.
16. Schweitzer BI, Dicker AP, Bertino JR. Dihydrofolate reductase as a therapeutic target. FASEB J 1990; 4:2441–2452.
17. Brown JD, Warren PV. Antibiotic discovery: is it all in the genes? Drug Discov Today 1998; 3: 564–566.

18. Read TD, Gill SR, Tettelin H, Dougherty BA. Finding drug targets in microbial genomes. Drug Discov Today 2001; 6:887–892.
19. Chalker A, Lunsford R. Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. Pharmacol Ther 2002; 95:1.
20. Wolf YI, Aravind L, Koonin EV. *Rickettsiae* and *Chlamydiae*: evidence of horizontal gene transfer and gene exchange. Trends Genet 1999; 15:173–175.
21. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
22. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999; 27: 260–262.
23. Haft DH, Loftus BJ, Richardson DL, et al. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 2001; 29:41–43.
24. Zhang Z, Schaffer AA, Miller W, et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 1998; 26:3986–3990.
25. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 278:631–637.
26. Galperin MY, Walker DR, Koonin EV. Analogous enzymes: independent inventions in enzyme evolution. Genome Res 1998; 8:779–790.
27. Perriere G, Duret L, Gouy M. HOBACGEN: database system for comparative genomics in bacteria. Genome Res 2000; 10:379–385.
28. Bruccoleri RE, Dougherty TJ, Davison DB. Concordance analysis of microbial genomes. Nucleic Acids Res 1998; 26:4482–4486.
29. Domagala JM, Hanna LD, Heifetz CL, et al. New structure-activity relationships of the quinolone antibacterials using the target enzyme. The development and application of a DNA gyrase assay. J Med Chem 1986; 29:394–404.
30. Schonbrunn E, Sack S, Eschenburg S, et al. Crystal structure of UDP-*N*-acetylglucosamine enolpyruvyltransferase, the target of the antibiotic fosfomycin. Structure 1996; 4:1065–1075.
31. Chalker AF, Minehart HW, Hughes NJ, et al. Systematic identification of selective essential genes in *Helicobacter pylori* by genome prioritization and allelic replacement mutagenesis. J Bacteriol 2001; 183:1259–1268.
32. Zhou H, Ranish JA, Watts JD, Aebersold R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. Nat Biotechnol 2002; 20:512–515.
33. Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 2001; 19:242–247.
34. Cagney G, Emili A. *De novo* peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. Nat Biotechnol 2002; 20:163–170.
35. Schmid MB, Kapur N, Isaacson DR, Lindroos P, Sharpe C. Genetic analysis of temperature-sensitive lethal mutants of *Salmonella typhimurium*. Genetics 1989; 123:625–633.
36. Xia M, Lunsford RD, McDevitt D, Iordanescu S. Rapid method for the identification of essential genes in *Staphylococcus aureus*. Plasmid 1999; 42:144–1449.
37. Thanassi JA, Hartman-Neumann SL, Dougherty TJ, Dougherty BA, Pucci MJ. Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. Nucleic Acids Res 2002; 30:3152–3162.
38. Wilding EI, Brown JR, Bryant AP, et al. Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in Gram-positive cocci. J Bacteriol 2000; 182:4319–4327.
39. Russell RB, Eggleston DS. New roles for structure in biology and drug discovery. Nat Struct

Biol 2000; 7(Suppl):928–930.

40. Gilliland GL, Teplyakov A, Obmolova G, et al. Assisting functional assignment for hypothetical *Haemophilus influenzae* gene products through structural genomics. Curr Drug Targets Infect Disord 2002; 2:339–353.
41. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res 2000; 28: 235–242.
42. Chambers SP. High-throughput protein expression for the post-genomic era. Drug Discov Today 2002; 7:759–765.
43. Blundell TL, Jhoti H, Abell C. High-throughput crystallography for lead discovery in drug design. Nature Rev Drug Disc 2002; 1:45–54.
44. Renfrey S, Featherstone J. Structural proteomics. Nature Rev Drug Disc 2002; 1:175–176.
45. Yee A, Chang X, Pineda-Lucena A, et al. An NMR approach to structural proteomics. Proc Natl Acad Sci USA 2002; 99:1825–1830.
46. Berman HM, Bhat TN, Bourne PE, et al. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 2000; 7(Suppl):957–959.
47. Gerstein M. Integrative database analysis in structural genomics. Nat Struct Biol 2000; 7(Suppl): 960–963.
48. Edwards AM, Arrowsmith CH, Christendat D, et al. Protein production: feeding the crystallographers and NMR spectroscopists. Nat Struct Biol 2000; 7(Suppl):970–972.
49. Buchanan SG, Sauder JM, Harris T. The promise of structural genomics in the discovery of new antimicrobial agents. Curr Pharm Des 2002; 8:1173–1788.
50. Lesley SA. High-throughput proteomics: protein expression and purification in the postgenomic world. Prot Expr Purif 2001; 22:159–164.
51. Pfuetzner RA, Bochkarev A, Frappier L, Edwards AM. Replication protein A. Characterization and crystallization of the DNA binding domain. J Biol Chem 1997; 272:430–434.
52. Udwary D, Merski M, Townsend C. A method for prediction of the locations of linker regions within large multifunctional proteins, and application to a type I polyketide synthase. J Mol Biol 2002; 323:585–598.
53. Elofsson A, Sonnhammer EL. A comparison of sequence and structure protein domain families as a basis for structural genomics. Bioinformatics 1999; 15:480–500.
54. Ostermeier C, Michel H. Crystallization of membrane proteins. Curr Opin Struct Biol 1997; 7: 697–701.
55. Chiu ML, Nollert P, Loewen MC, et al. Crystallization *in cubo*: general applicability to membrane proteins. Acta Crystallogr D Biol Crystallogr 2000; 56:781–784.
56. Capdeville R, Buchdunger E, Zimmermann J, Matter A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. Nature Rev Drug Disc 2002; 1:493–502.
57. Betz SF, Baxter SM, Fetrow JS. Function first: a powerful approach to post-genomic drug discovery. Drug Discov Today 2002; 7:865–871.
58. Baxter SM, Fetrow JS. Sequence- and structure-based protein function prediction from genomic information. Curr Opin Drug Discov Dev 2001; 4:291–295.
59. Davies C, Heath RJ, White SW, Rock CO. The 1.8 A crystal structure and active-site architecture of beta-ketoacyl-acyl carrier protein synthase III (FabH) from *Escherichia coli*. Structure Fold Des 2000; 8:185–195.
60. Colovos C, Cascio D, Yeates TO. The 1.8 A crystal structure of the ycaC gene product from *Escherichia coli* reveals an octameric hydrolase of unknown specificity. Structure 1998; 6:1329–1337.
61. Cort JR, Yee A, Edwards AM, Arrowsmith CH, Kennedy MA. Structure-based functional classification of hypothetical protein MTH538 from *Methanobacterium thermoautotrophicum*. J Mol Biol 2000; 302:189–203.

62. Minasov G, Teplova M, Stewart GC, Koonin EV, Anderson WF, Egli M. Functional implications from crystal structures of the conserved *Bacillus subtilis* protein Maf with and without dUTP. Proc Natl Acad Sci USA 2000; 97:6328–6333.
63. Teplova M, Tereshko V, Sanishvili R, et al. The structure of the *yrdC* gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. Protein Sci 2000; 9: 2557–2566.
64. Lewis HA, Furlong EB, Laubert B, et al. A structural genomics approach to the study of quorum sensing: crystal structures of three LuxS orthologs. Structure 2001; 9:527–537.
65. Buchan DW, Shepherd AJ, Lee D, et al. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res 2002; 12:503–514.
66. Pearl FM, Lee D, Bray JE, Buchan DW, Shepherd AJ, Orengo CA. The CATH extended protein-family database: providing structural annotations for genome sequences. Protein Sci 2002; 11:233–244.
67. Williams MG, Shirai H, Shi J, et al. Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. Proteins 2001; Suppl 5:92–97.
68. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001; 310:243–257.
69. Liu Y, Luscombe NM, Alexandrov V, et al. Structural genomics: a new era for pharmaceutical research. Genome Biol 3: REPORTS4004. 2002.
70. Demain AL. Microbial natural products: alive and well in 1998. Nat Biotechnol 1998; 16: 3–4.
71. Farnet CM, Staffa A, Zazopoulos E. High throughput method for discovery of gene clusters. Canadian Appl. CA 2,352,451. Ecopia Biosciences, Inc., Canada. 2001.
72. Farnet CM, Mercure S, Nowacki P, Staffa A, Zazopoulos E. Gene cluster for everinomicin biosynthesis. PCT Int. Appl. WO 0155180. Ecopia Biosciences, Inc., 2001.
73. Farnet CM, Staffa A, Zazopoulos E. Gene cluster for ramoplanin biosynthesis. PCT Int. Appl. WO 0231155. Ecopia Biosciences, Inc., 2002.
74. Farnet CM, Staffa A, Yang X. Gene and proteins for rosaramicin biosynthesis. Canadian Appl. CA 2,391,131. Ecopia Biosciences, Inc., Canada, 2002.
75. Farnet CM, Staffa A. Genes and proteins for the biosynthesis of anthramycin. Canadian Appl. CA 2,386,587. Ecopia Biosciences, Inc, Canada, 2002.
76. Ahlert J, Shepard E, Lomovskaya N, et al. The calicheamicin gene cluster and its iterative type I enediyne PKS. Science 2002; 297:1173–1176.
77. Newman DJ, Cragg GM, Snader KM. The influence of natural products upon drug discovery. Nat Prod Rep 2000; 17:215–234.
78. Rondon MR, August PR, Bettermann AD, et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 2000; 66:2541–2547.
79. Brady SF, Chao CJ, Clardy J. New natural product families from an environmental DNA (eDNA) gene cluster. J Am Chem Soc 2002; 124:9968–9969.
80. Wang GY, Graziani E, Waters B, et al. Novel natural products from soil DNA libraries in a streptomycete host. Org Lett 2000; 2:2401–2404.
81. Brady SF, Chao CJ, Handelsman J, Clardy J. Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. Org Lett 2001; 3:1981–1984.
82. Courtois S, Cappellano CM, Ball M, et al. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl Environ Microbiol 2003; 69:49–55.

83. Jensen PR, Fenical W. Marine microorganisms and drug discovery: Current status and future potential. In: Fusetani N (ed). Drugs from the Sea. Basel: Karger, 2000, pp. 6–29.
84. Faulkner DJ, Harper MK, Haygood MG, Salomon CE, Schmidt EW. Symbiotic bacteria in sponges: source of bioactive substances. In: Fusetani N (ed). Drugs from the Sea. Basel: Karger, 2000, pp. 107–119.
85. Ness JE, Welch M, Giver L, et al. DNA shuffling of subgenomic sequences of subtilisin. Nat Biotechnol 1999; 17:893–896.
86. Stemmer WP. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc Natl Acad Sci USA 1994; 91:10,747–10,751.
87. Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WP, del Cardayre SB. Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature 2002; 415:644–646.
88. Patnaik R, Louie S, Gavrilovic V, et al. Genome shuffling of *Lactobacillus* for improved acid tolerance. Nat Biotechnol 2002; 20:707–712.

27

基因组法开发疫苗

Rino Rappuoli, Vega Masignani, Mariagrazia Pizza, Guido Grandi and John L. Telford

引言

由基因组研究所（The Institute for Genomic Research）维护的微生物数据库，列出了 140 多种细菌的全基因组，并且全世界不同实验室对 300 多个微生物正在测序。几年前还不可思议的大量信息以及尖端计算工具的迅速发展，已改变了对原核世界的理解，并将影响今后的微生物研究。科研人员利用越来越大的数据库和特殊工具，只根据序列分析，而不依赖传统的费力、昂贵、费时的生化方法，便可快速推断出蛋白质功能。

对病原细菌测序的根本目的是理解感染疾病的过程，由此能开发分子诊断探针，确定新药物靶标和采取预防措施，处理微生物引起的感染，在这个意义上，生物信息学最有前途的应用是疫苗领域。实际上，细菌基因组的全序列，能够从完全不同的角度提供开发疫苗的机会。在一个基因组内，所有蛋白质抗原都是一样可见的，与它们的表达量和可检测方式（体内、体外或生长的某个阶段）无关，不仅可以用传统的生化、血清和微生物方法筛选的抗原进行鉴定，而且还可能发现新抗原[1]。

表 1　研究基因的计算机程序

程　序	网　址	适用范围
BLAST/PSI-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	同源搜索
FASTA	GCG Package, in house	同源搜索
PSORT	http://psort.ncbi.ac.jp/	信号肽、跨膜片段和一般定位预测
SignalP	http://www.cbc.dtu.dk/services/SignalP/	信号肽预测
SPScan	GCG Package, in house	信号肽预测
TMpred	http://www.ch.embet.org/software/TMPRED-form.html	跨膜蛋白和方位预测
TopPred2	http://bioweb.pasteur.fr/sequanal/internal	疏水片段和膜蛋白拓扑学
Motifs	GCG Package, in house	已知蛋白模体
FindPattens	GCG Package, in house	用户界定蛋白模体
InterPro	http://www.ebi.ac.uk/interpro/	特征鉴定和蛋白家簇 A 整合资源
PredictProtein	http://www.embl-heidelberg.de/predictprotein	结构预测
PSIPRED	http://bioinf.ucl.ac.uk/psipred/	结构预测

尖端计算机软件可以预测基因产品的功能，寻找与其他病原菌产生已知毒性因子的

同源性，预测新识别可读框（ORF）的细胞位置（表 1），就可能通过计算机模拟（*in silico*）分析，寻找细菌病原菌潜在保护性抗原，然后在保护性免疫模型中进行测验，这种方法，命名为反向疫苗学（reverse vaccinology），已经用于寻找抗 B 型脑膜炎奈瑟氏球菌（*Neisseria meningitidis*）的新候选疫苗[2]，正用于开发抗其他病原菌的疫苗[3]。本章介绍基本方法和应用。

从基因组到抗原：一个新的范例

用"计算机模拟"寻找候选疫苗

细菌蛋白作为抗原的最基本条件是在细胞中的部位，胞内蛋白质不可能是免疫目标，而细胞的表面结构和分泌物，更容易接触抗体——抗细菌病原菌最基本免疫效应分子。图 1 总结了当作疫苗目标的蛋白质类型，细菌有一些控制新合成蛋白进入胞外系统，因此，与寄主相互作用的胞外酶和蛋白（如黏附和毒性因子）表面表达有关。虽然，在革兰氏阴性菌和革兰氏阳性菌中有些系统相同，另一些系统则有特异性。

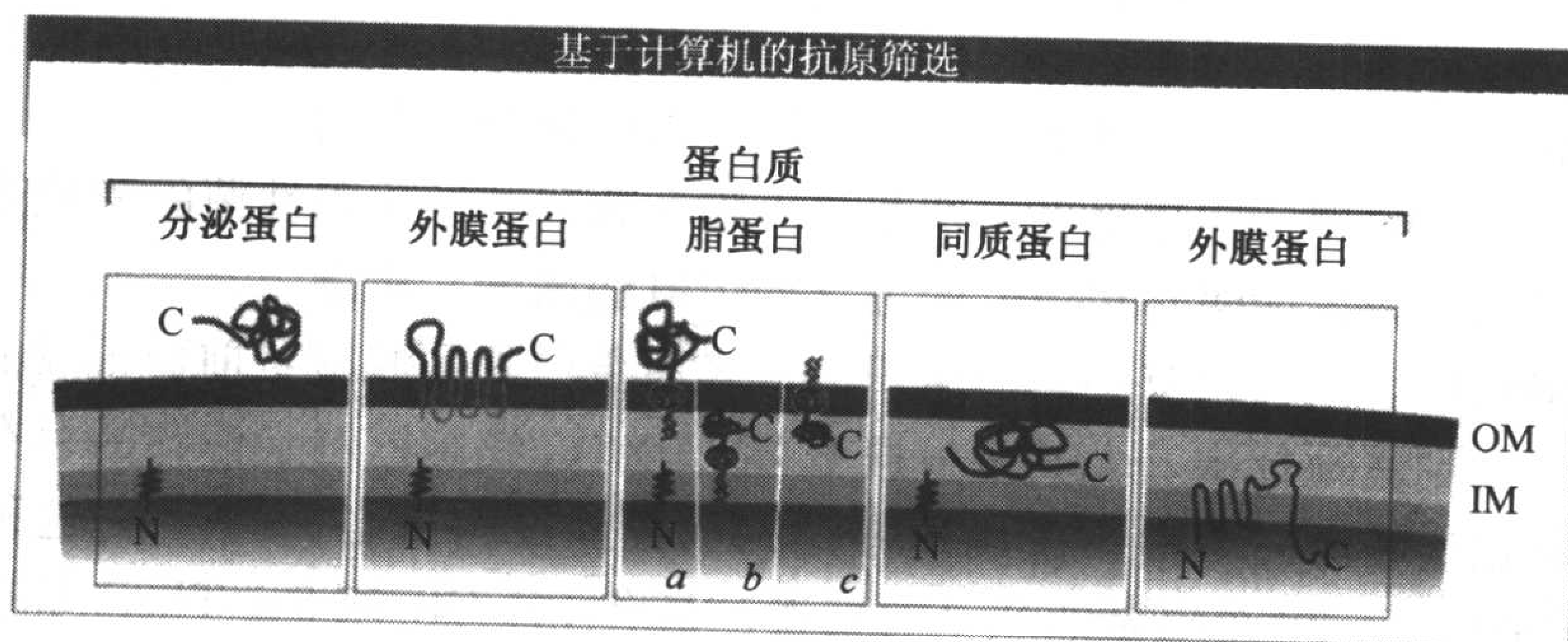


图 1 能作为抗原的蛋白类型。从左到右：分泌蛋白、外膜（OM）蛋白、脂蛋白、周质蛋白、内膜（IM）蛋白。

无论是哪类菌，大多数分泌蛋白作为前体而合成，并在氨基端携带"邮政编码"——信号肽，对信号肽的识别，导致新生蛋白进入普通分泌途径[4]，在通过内膜传输过程中，信号肽被特殊的信号肽酶切断。虽然不同氨基端信号肽的初级结构有很少相似性，却都有三个保守区域：带正电荷 N 区、疏水核心和切点（如 Ala-X-Ala）之前不带电荷的极性 C 区[5]，前导肽的结构特征可能因蛋白分泌方式不同而稍有区别。

例如，脂蛋白信号肽有一个相当保守的脂框（lipobox），该脂框总含有一个半胱氨酸残基（一般以亮氨酸-X-X-半胱氨酸形式存在），在前体切开前，该半胱氨酸被二酰甘油转移酶进行脂修饰[6]。当脂蛋白横穿细胞膜后，脂蛋白通过氨基末端的脂修饰半胱氨酸残基而锚定在内膜或外膜上。双精氨酸转位装置系统（TAT 系统），是一个特异不依赖于普通分泌途径的分泌系统，最初在植物中发现，也见于一些革兰氏阳性菌和阴性菌中[7]。该途径的底物特征是：前导序列的疏水核心前有一致序列 Arg-Arg-X-Phe-Leu-Lys，包含一对连续的精氨酸残基。

在对这些特征充分认识的基础上，开发了一些计算机程序，根据氨基末端序列预测

分泌蛋白，几个软件产品（见表1）能在整个细菌基因组范围内快速自动地识别分泌蛋白。虽然，信号肽识别能更有效地预测分泌蛋白，但是，表面暴露的其他特征可以支持这种预测，并用于发现那些缺乏典型信号肽的表面蛋白。

可以用几种不同方式获得复合蛋白跨越革兰氏阴性细菌外膜的转位，最简单的是所谓自转运分泌机制（autotransporter secretion mechanism）[8]，它是在研究淋病奈瑟氏球菌（*Neisseria gonorrhoeae*）IgA1蛋白酶时首次被揭示[9]。自身转位的蛋白质通过C端跨越外膜而输出，这种C端区域是以反向平行双亲性β折叠组成的孔状式样定位于外膜内，该蛋白的C端残基都是苯丙氨酸或色氨酸，这是把蛋白质锚定到外膜上所必需的，在该末端氨基酸残基之前的序列，由带电荷/极性以及芳香族/疏水的氨基酸交替组成，形成了(Y，F，W，L，I，V)-X-(F，W)式的明显特征。

跨越内膜和外膜更复杂的输出机制，包括近期发现的类型III和类型IV[10, 11]，这些系统涉及不同数量的组分，这些组分聚集成跨越内外膜的大型结构，允许特殊因子直接分泌到细胞外，或者分泌到宿主细胞膜内。这些系统的原型是耶尔森菌的*Yop*系统和根癌土壤杆菌的VirB系统[13]，直接识别这些系统很不容易，因为通过类型III和类型IV机制输出的蛋白质没有特征性序列或结构，而在不同组分中序列相似水平很低。然而，编码分泌系统的基因通常一起被转录，或者最起码在基因组上连续排列，包括几种腺苷三磷酸结合蛋白、膜镶嵌蛋白、细胞周质蛋白和分泌蛋白。

对疏水片段的预测，广泛用于识别革兰氏阳性菌和革兰氏阳性菌的跨膜蛋白，跨膜蛋白常涉及运输（如渗透酶）和信号传导（如组氨酸激酶）机制。

革兰氏阳性菌表面结构有不同的组织，对表面蛋白的预测除了前面提及的一般输出途径外，还涉及到其他标准。尤其是根据C端LPXTG细胞壁锚定修饰的存在来预测表面蛋白。这种锚定修饰对蛋白质正确地锚定到肽聚糖结构是必需的[14]。这种氨基酸模型位于距C端约25～30个氨基酸残基的位置，富含Pro-Gly或Ser-Thr，紧接横跨内膜疏水片段，最后是带正电荷的一个短尾巴。其他革兰氏阳性菌的特殊表面结构，包括通过疏水作用或带电荷区域锚定的一群蛋白质，以及通过重复序列结合脂磷壁酸蛋白。最后，可以通过与已知毒性因子的同源性或通过与其他微生物已知表面蛋白的同源性，来寻找革兰氏阳性菌和阴性菌的候选疫苗。

高通量表达

用上述标准寻找候选疫苗可能要筛选大量基因，覆盖基因组中总数达25%的可读框。为了产生与这些基因相对应的每一重组蛋白，必须用简单的方法来克隆和表达大量基因，幸运的是，机器人的发展和PCR反应使这成为现实。

根据基因组序列，可以设计PCR多核苷酸序列，每一对多核苷酸应包含与表达载体的克隆位点相对应的限制性酶切位点，或者包含重组酶的识别位点以便用体外重组构建质粒。每一个PCR反应产物然后克隆到两个单独的表达载体，这些载体要么包含编码由连续六个组氨酸残基组成的标记序列（6×His），要么包含编码谷胱苷肽转移酶（GST）的序列，这些标记序列可以通过简单的柱层析法快速纯化重组蛋白。

抗原检测

反向疫苗学方法的关键是快速检测分子对抗病原菌的免疫保护，最简单的方法是用

重组抗原免疫老鼠，获取免疫血清，然后用酶联免疫分析（ELISA）或流式细胞仪来检测它结合细菌表面天然抗原的能力。这种方法虽快捷，但对表面抗原的识别不一定代表免疫保护，与简单的表面识别实验相比，这种在特殊抗体存在下对革兰氏阴性菌补体介导的裂解实验繁琐些，但在一些情况下，这种杀菌活力与对人体免疫保护有很好的相关性。革兰氏阳性菌，经常在新分离嗜中性白细胞的存在下，进行抗体和补体依赖性调理吞噬作用（opsonophagocytosis）分析，不过这比直接杀菌检测法复杂多了。然而，在某些情况下，筛选保护性抗原的唯一办法是使用动物感染模型。

抗 B 型脑膜炎奈瑟氏球菌疫苗

寻找抗原

随着抗流感嗜血菌（*Haemophilus influenzae*）缀合糖疫苗（glycoconjugate vaccine）1988 年成功用于临床实践以来，同样方法用来试制抗肺炎链球菌和脑膜炎奈瑟氏球菌的类似产品[15]。虽然针对肺炎链球菌和脑膜炎奈瑟氏球菌血清型 A、C、Y 和 W135 以多糖为基础的疫苗，已可使用或处于开发最后阶段，但是，B 型脑膜炎奈瑟氏球菌是主要挑战对象，针对荚膜的传统疫苗不能对付这个血清型，主要因为 B 型多糖的特殊结构类似人脑组织一个组分的结构，这样，B 型多糖在人体中的免疫原性很低，尝试突破这个极限可能会导致自身免疫反应[17]。鉴于此，最近的策略主要是寻找有免疫原性的蛋白质分子，而不是多糖分子[18]，然而，尽管许多研究组经过 40 年的努力，所有传统的生化和微生物方法，都不能产生抗 B 型脑膜炎奈瑟氏球菌的通用疫苗。

采用计算机新技术和高致病力 B 型脑膜炎奈瑟氏球菌血清菌株的完整基因组序列[19]，用反向疫苗学方法开发抗 B 型脑膜炎奈瑟氏球菌的候选疫苗[2]，图 2 总结了脑膜炎奈瑟氏球菌抗原筛选的一般策略。通过计算机模拟筛选，发现 570 个新可读框，预计它们编码分泌蛋白或表面暴露蛋白，因此代表新的潜在候选疫苗。筛选的基因产物约有一半与已知蛋白同源，然而，其他的为保守假定蛋白（与其他生物假定蛋白极为相似，但功能未知）和假定蛋白（在数据库中没有同源物，可能为种属特异蛋白）。绝大多数推测蛋白是膜镶嵌蛋白，其次是周质蛋白、脂蛋白、外膜蛋白和分泌蛋白，它们不足总数的 15%。

在 570 个选择的可读框中，350 个成功在大肠杆菌中以 6×His 或 GST 融合蛋白成功克隆和表达，预计不能表达的可读框绝大多数编码两个跨膜区域以上的蛋白。这些蛋白特别难于在大肠杆菌中表达，即使表达，也通常对细胞有毒，然后表达蛋白通过镍螯合树脂或谷胱苷肽聚合树脂的单柱层析法纯化，其中 344 个表达蛋白用于免疫四个 CD1 小鼠，获得的血清用细菌细胞裂解物的免疫印迹法分析，确定该蛋白是否在脑膜炎奈瑟氏球菌中表达，如果表达，则对细菌全细胞进行酶联免疫吸附分析（ELISA）、间接免疫荧光和流式细胞仪分析，确定是否可在细菌表面检测到，这些实验找出了 91 个新表面暴露蛋白。

为了分析这些候选疫苗作为抗原的潜力，检测抗重组蛋白的免疫血清，看其补体依赖性杀菌能力。杀菌实验中的高价血清与人类抵抗疫病很好相关，在临床试验中，这种血清已被很多管理部门批准为免疫保护的替代品。这些分析发现了 29 个新抗原，这些

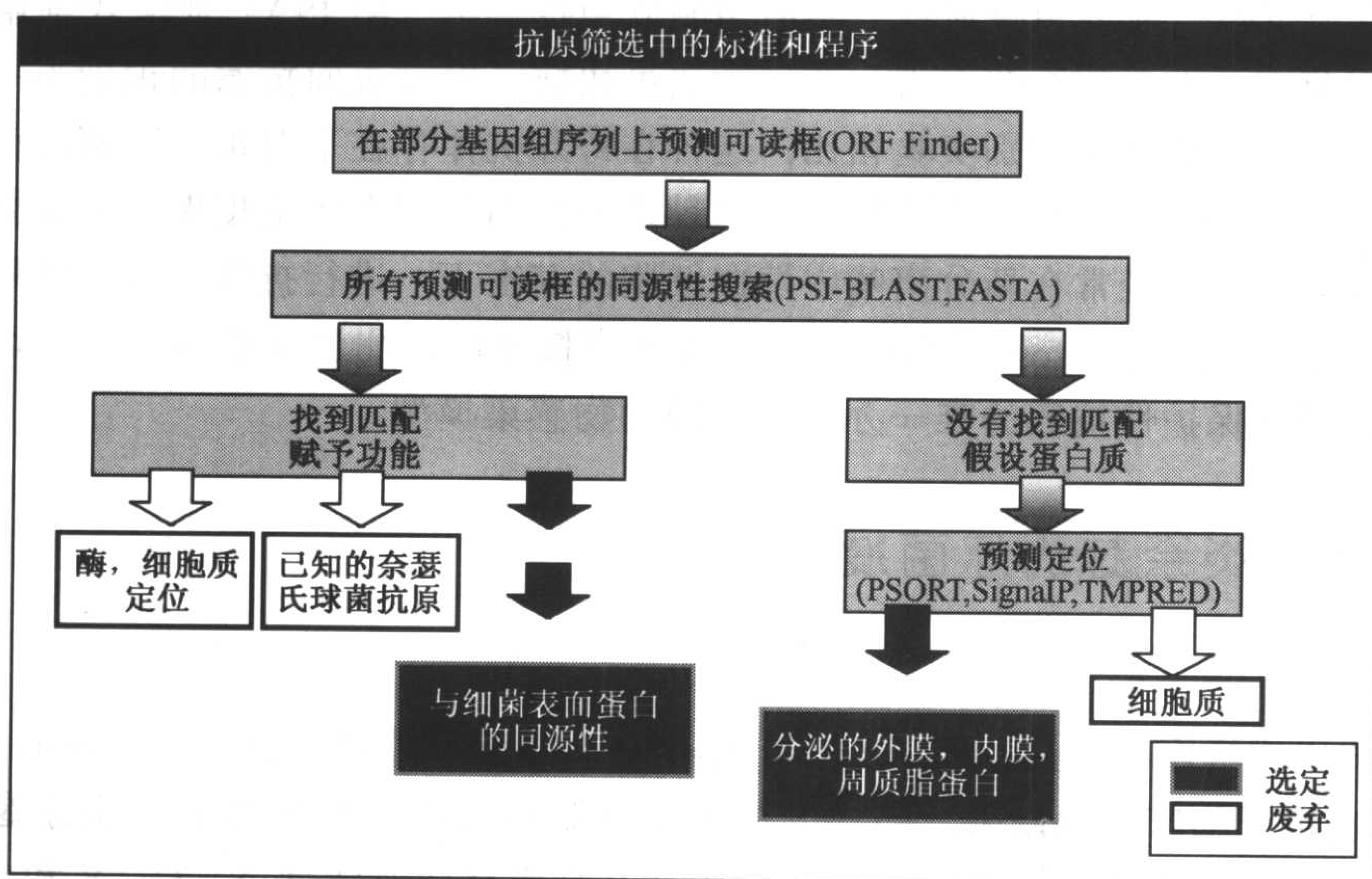


图 2　计算机模拟抗原筛选策略的主要步骤及计算机程序。

抗原能诱导高价杀菌活性的血清。这样，在两年左右的时间里，筛选到的候选疫苗比过去 40 多年对脑膜炎奈瑟氏球菌得到的候选疫苗还要多，抗原筛选结果见图 3。

抗原变异性和交叉保护

许多已知表面暴露抗原的氨基酸序列可变，其抗原性也是可变的。实际上，以包含主要表面抗原外膜为基础开发的疫苗，显示出对原始菌株感染的良好保护，却对其他流行菌株没有保护作用[20]）。因此，评估基因组筛选中获得抗原的变异性，以及它们诱导广泛免疫保护的能力非常重要。

为做到这一点，根据表面暴露程度和杀菌效价选择了 7 个抗原，并由 PCR 和 DNA 印迹杂交，确定相应基因在实验菌株中的存在，这些菌株是全世界主要致病脑膜炎奈瑟氏球菌的分离株，包括血清型 A、C、Y、Z 和 W135。编码 7 种抗原的基因在所有实验菌株中都存在，有一些抗原基因也在乳糖奈瑟氏球菌（*Neisseria lactamica*）、灰色奈瑟氏球菌（*Neisseria cinerea*）或淋病奈瑟氏球菌（*Neisseria gonorrhoeae*）中被发现。对这些菌株的基因测序证明，5 种抗原高度保守，2 个基因在蛋白的某些区域发生了变异，抗血清对杀菌实验中的一组细菌有交叉保护。

这样，在极短时间内已发现了一些在临床前试验中效果很好的一些新蛋白抗原。这些抗原将进入人体临床试验，以确定是否可以抗这种重要医学病原菌。

革兰氏阳性菌试验

B 群链球菌

为了证明反向疫苗方法的通用性（图 3），决定将其应用到抗革兰氏阳性人类病原菌 B 群链球菌（无乳链球菌 *Streptococcus agalactiae*）的疫苗设计上，该病原菌是发达

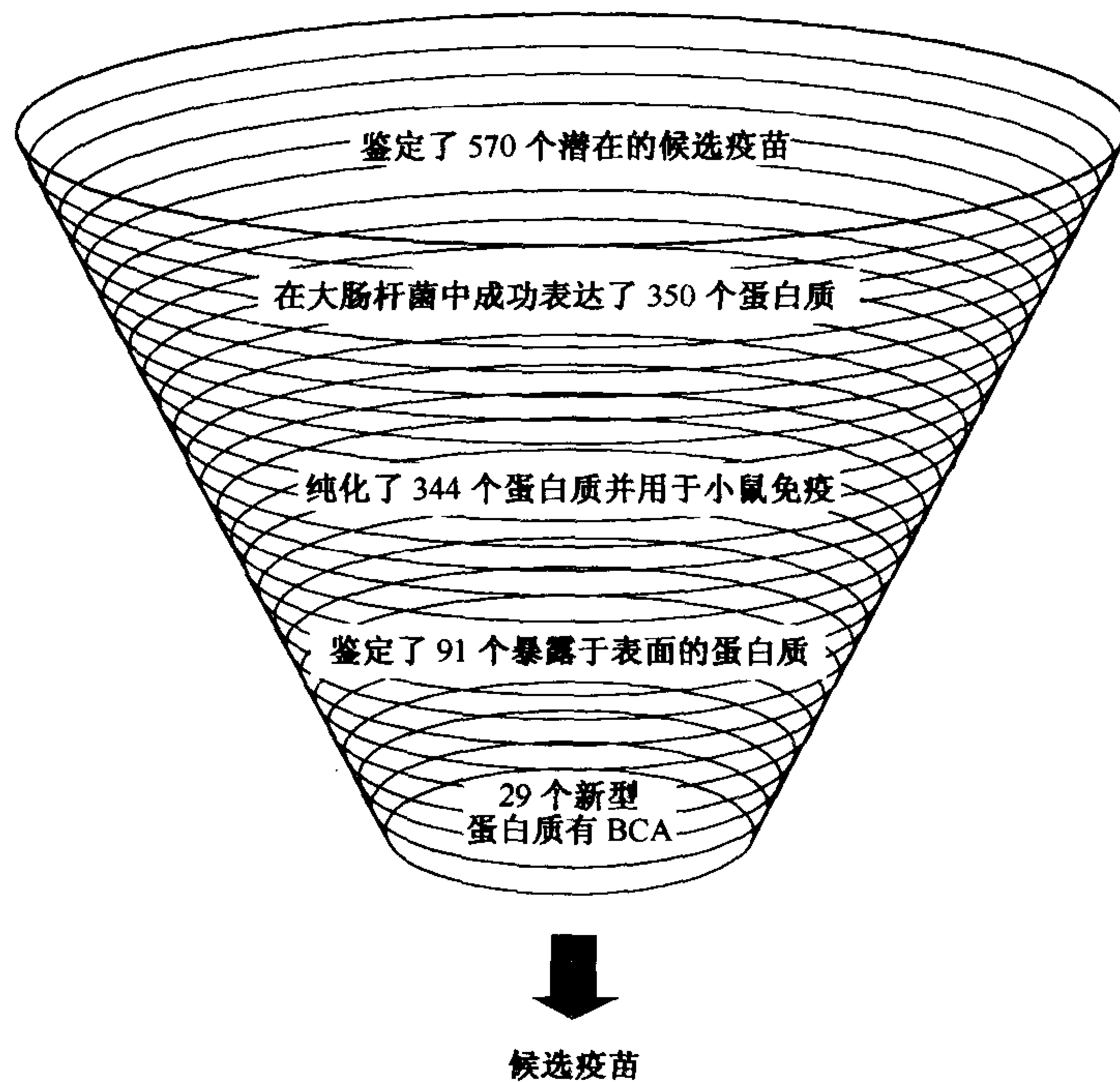


图 3 反向疫苗方法获得抗 B 型脑膜炎奈瑟氏球菌疫苗的结果

国家新生儿败血症的重要诱因[21]。这种细菌是在分娩时，由带菌母亲传给婴儿，通常在分娩 24 小时内导致灾难性菌血症和死亡。给母亲注射疫苗将诱导抗该细菌的抗体，这些抗体能在婴儿出生前通过胎盘传给婴儿，并保护其免受侵袭性感染。这种母亲的保护方式已在小鼠模型中得到验证[23]，并显示，如果母亲已有抗细菌的高效价抗体，婴儿很少被感染[24]。

至今，动物模型中最佳保护抗原是细菌多糖荚膜[22]，遗憾的是，起码有 9 种不同荚膜血清型，在这些血清型之间很少甚至没有交叉保护。至今对不同血清型内和不同血清型之间的基因组变异知之甚少，尚不清楚是否可以发现对流行菌株有交叉保护的足够保守蛋白抗原，因此，对开发这种革兰氏阳性菌疫苗所存在的问题，在某种程度上不同于在寻找抗脑膜炎奈瑟氏球菌疫苗所面临的问题。

完整 B 群链球菌基因组

在与基因组研究所的合作中，已确定和分析了乳糖链球菌血清型 V 菌株的完整基因组[25]，该基因组可编码 2175 个可读框，其中 650 个暴露于细菌表面，已成功在大肠杆菌中表达了约 350 个可读框，并将其用于免疫小鼠。在酶联免疫吸附分析（ELISA）和流式细胞仪分析中，用血清抗完整细菌，证明了 55 种抗原确实在细菌表面表达，这些抗原正通过体内体外模型，评估其抗 B 型链球菌侵袭性感染的能力。因为对该菌的变异性知之甚少，只能在基因组水平和单个基因水平评估基因的变异性。

基因组水平的血清型变异

已经采用完整的基因组杂交，检测代表多血清型 B 群链球菌的 19 个菌株所有基因的存在或缺失。由 PCR 反应合成代表测序菌株中所有检测可读框的短序列，并将 PCR 产物排列在基因芯片上，然后与 19 个菌株中每个菌株的标记 DNA 杂交，这些杂交信号与参考菌株基因组芯片杂交信号进行比较，某一杂交信号的缺失表明，该菌株要么缺少该基因，要么该基因已进化到高一级程度。

在参考菌株中共有 401 个可读框，至少与一个实验菌株无杂交信号，表明在这个实验菌株中，这些可读框要么缺失，要么有高度多样性。在这些基因中，发现 90% 存在 15 个基因簇中，每个基因簇最少由 5 个邻接基因组成，在某些情况下，这些基因簇有原噬菌体的明显特征，或两侧有转座子序列，此外，其中 10 个区域核苷酸组分与基因组的其他区域不同，表明它们是在参考菌株中通过未知 DNA 的水平转移而获得。只部分菌株存在的基因中，发现 37 个零散随机分布在基因组中，显然，好候选疫苗必须来自大多数流行菌株都有的那些基因，有趣的是，基因的有无与血清型没有明显关系。

基因水平的变异性

为了评价不同血清型之间单个蛋白质在氨基酸序列水平的变异性，根据预测，从 19 个菌株中选择了 8 个细胞质管家基因和 11 个表面蛋白基因，编码已知主要表面蛋白的基因不包括在这次分析中，因为这些基因的大多数高度变异，可能是由于宿主免疫系统的压力。令人惊奇的是，所有受测基因都高度保守，不论相应蛋白位于胞质中，还是位于细菌表面。

一般情况下，在测试菌株中预测的蛋白质在氨基酸序列上的一致性超过 97%，还不清楚，为什么其表面暴露蛋白比其他侵袭性病原菌（例如脑膜炎奈瑟氏球菌）变异得更少，这反映了 B 群链球菌在肛门和阴道区域生长繁殖的事实，这些部位不是特别活跃的免疫位点。无论什么原因，基因保守性预示了发现能预防侵袭疾病的交叉血清抗原的可能性。

有趣的是，根据核苷酸序列的种系分析表明，虽然在一定程度上菌株按血清类型聚成一簇，但是一些菌株却与其他血清型菌株聚成一簇。结合全基因组杂交实验，这些数据表明实际遗传谱系不依赖于血清型，并暗示血清型变换可能在 B 群链球菌中相对频繁。这也不奇怪，因为基因组杂交表明，基因组 DNA 有大量流动性。

反向疫苗学的未来

短时间内通过对抗 B 型脑膜炎奈瑟氏球菌的几个新保护抗原的发现，证明反向疫苗学的基本概念有效[2]，这些抗原发现的速度证明该方法的有效性。而用基因组方法对 B 群链球菌研究的初期结果，导致对大量在细菌表面表达新高度保守抗原的发现[25]，极有可能在它们中间发现可作为疫苗的保护性抗原。因此，该方法对大范围病原菌一般是适用的，唯一限制要有基因组序列和检测抗原诱导保护性免疫反应能力的适当体内外模型。该方法的明显优点是，通过保护检测所有抗原，已在大肠杆菌中作为溶解重组蛋

白而生产，这样，候选抗原将来可以直接用于大规模工业化生产。

然而，结合其他基因组学技术，反向疫苗法可以得到进一步完善。代表基因组内所有可读框的 DNA 微阵列，可以与体外培养或从感染动物甚至感染的病人中分离的细菌中抽提的 RNA 杂交。产生的数据可用来发现那些在细菌中大量表达，并将导致有效免疫反应的基因。另外，对人类疾病模型中体内表达基因的鉴定，将进一步帮助精炼筛选程序，我们已从黏附生长在上皮细胞的细菌中抽提的 RNA 进行微阵列杂交，发现抗 B 型脑膜炎奈瑟氏球菌的新保护性抗原，证明这个策略有效。而在直接基因组方法中，我们未能获得这个抗原[26]。

蛋白质组学方法也能帮助精炼抗原筛选方法，采用双向电泳和质谱法分析沙眼衣原体（*Chlamydia trachomatis*）表面相关蛋白，发现了抗此感染的大量潜在保护性抗原[27]，所用的新技术现在能快速进行这些实验，这些新技术能弥补传统的蛋白质组策略（见参考文献[28]中的综述）。

最后，虽然反向疫苗学只能用于细菌病原菌，在理论上也可用于病毒或真核生物的寄生物中，不足之处是基因组大小和检测重组抗原需要合适的模型。可以肯定地说，全基因组序列信息导致了现代疫苗研究的革命。

（汪世山 译）

参考文献

1. Rappuoli R. Reverse vaccinology. Curr Opin Microbiol 2000; 3:445–450.
2. Pizza M, Scarlato V, Masignani V, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000; 287:1816–1820.
3. Wizemann TM, Heinrichs JH, Adamou JE, et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect Immun 2001; 69:1593–1598.
4. Economou A. Following the leader: bacterial protein export through the Sec pathway. Trends Microbiol 1999; 7:315–320.
5. von Heijne G. The signal peptide. J Membr Biol 1990; 115:195–201.
6. von Heijne G. The structure of signal peptides from bacterial lipoproteins. Protein Eng 1989; 2:531–534.
7. Berks BC, Sargent F, De Leeuw E, et al. A novel protein transport system involved in the biogenesis of bacterial electron transfer chains. Biochim Biophys Acta 2000; 1459:325–330.
8. Henderson IR, Navarro-Garcia F, Nataro JP. The great escape: structure and function of the autotransporter proteins. Trends Microbiol 1998; 6:370–378.
9. Pohlner J, Halter R, Beyreuther K, Meyer TF. Gene structure and extracellular secretion of *Neisseria gonorrhoeae* IgA protease. Nature 1987; 325:458–462.
10. Cornelis GR, Van Gijsegem F. Assembly and function of type III secretory systems. Annu Rev Microbiol 2000; 54:735–774.
11. Christie PJ. Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. Mol Microbiol 2001; 40:294–305.
12. Cornelis GR, Wolf-Watz H. The *Yersinia* Yop virulon: a bacterial system for subverting eukaryotic cells. Mol Microbiol 1997; 23:861–867.

13. Zupan JR, Ward D, Zambryski P. Assembly of the VirB transport complex for DNA transfer from *Agrobacterium tumefaciens* to plant cells. Curr Opin Microbiol 1998; 1:649–655.
14. Navarre WW, Schneewind O. Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol Mol Biol Rev 1999; 63:174–229.
15. Lindberg AA. Glycoprotein conjugate vaccines. Vaccine 1999; 17:S28–S36.
16. Zollinger WD. New and improved vaccines against meningococcal disease. In: Levine MM, Woodrow GC, Cobon GS (eds). New Generation Vaccines. New York: Decker, 1997, pp. 468–488.
17. Hayrinen J, Jennings H, Raff HV, et al. Antibodies to polysialic acid and its *N*-propyl derivative: binding properties and interaction with human embryonal brain glycopeptides. J Infect Dis 1995; 171:1481–1490.
18. Jodar L, Feavers IM, Salisbury D, Granoff DM. Development of vaccines against meningococcal disease. Lancet 2002; 359:1499–1508.
19. Tettelin H, Saunders NJ, Heidelberg J, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 2000; 287:1809–1815.
20. Rosenstein NE, Fischer M, Tappero JW. Meningococcal vaccines. Infect Dis Clin North Am 2001; 15:155–169.
21. Schuchat A. Group B streptococcus. Lancet 1999; 353:51–56.
22. Baker CJ, Kasper DL. Group B streptococcal vaccines. Rev Infect Dis 1985; 7:458–467.
23. Paoletti LC, Wessels MR, Rodewald AK, Shroff AA, Jennings HJ, Kasper DL. Neonatal mouse protection against infection with multiple group B streptococcal (GBS) serotypes by maternal immunization with a tetravalent GBS polysaccharide-tetanus toxoid conjugate vaccine. Infect Immun 1994; 62:3236–3243.
24. Lin FY, Philips JB 3rd, Azimi PH, et al. Level of maternal antibody required to protect neonates against early-onset disease caused by group B streptococcus type Ia: a multicenter, seroepidemiology study. J Infect Dis 2001; 184:1022–1028.
25. Tettelin H, Masignani V, Cieslewicz MJ, et al. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci USA 2002; 99:12,391–12,396.
26. Grifantini R, Bartolini E, Muzzi A, et al. Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. Nat Biotechnol 2002; 20:914–921.
27. Montigiani S, Falugi F, Scarselli M, et al. Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. Infect Immun 2002; 70:368–379.
28. Grandi G. Antibacterial vaccine design using genomics and proteomics. Trends Biotechnol 2001; 19:181–188.

28

微生物蛋白质组学

Svend Birkelund, Brian B. Vandahl, Allan C. Shaw and Gunna Christiansen

引言

细菌蛋白质组作为基因组的补充

自从 1995 年第一个基因组发表以来[1]，已经测序了 140 种微生物基因组，还有 300 多种序列鉴定正在进行当中，这项工作正在由国家生物技术信息中心完成[2]。后基因组时代的焦点是功能基因组学，而蛋白质组学在其中发挥重要作用。基因组提供了某物种生物潜能的重要信息，但基因组是静态的，不能提供某个特定基因表达的信息，无法知道翻译后修饰的过程，也无法知道特定生物环境中某个蛋白质的调控。

活细胞是一个动态复杂的环境，其组成和结构无法通过基因组序列推测，细菌的每个种约有 30％预测蛋白质与公共数据库的记录没有同源性，或者与功能未知的蛋白质有同源性。转录分析揭示哪些基因转录成核糖核酸，运用微阵列技术大规模进行大量分析是可行的[3]（见第 22 章）。然而，mRNA 和蛋白质之间没有严格的关系[4,5]，而且也不知道翻译后修饰、加工和转运的情况，对细胞蛋白质总量的直接研究是蛋白质组学的任务。蛋白质组学定义为在某一特定条件下（如生长条件和特定时间），一种特定生态环境中经过翻译后修饰和加工的一整套蛋白质[6]。

技术

蛋白质组学，可以用很多不同方法进行研究，原则上包括两个步骤：第一，样品中蛋白质的分离。第二，蛋白质的鉴定。最普通的分离工具是双向电泳（2D PAGE），大规模鉴定蛋白质用质谱法，另外还有众所周知的方法，如，N 末端测定、免疫杂交、过量表达、斑点共定位（spot colocalization）、基因敲除等都可以用于蛋白质鉴定。

双向电泳分离蛋白

由于双向电泳具有高分辨率，该技术目前是微生物蛋白质组研究最得力的工具。第一向等电聚焦，在这个过程中，蛋白质根据各自等电点在 pH 不同梯度上分离，蛋白质停留在使其净电荷为零的 pH 处；第二向是蛋白质根据其分子质量的不同，SDS-PAGE 将蛋白质分开。得到胶的图像是由若干斑点构成，每个斑点代表一种蛋白质，横向和纵向分别代表等电点和相对分子质量。

样品制备

2D PAGE 的关键步骤是样品制备。没有哪一种样品制备方法可以通用于所有样品，因为，不同试剂对不同样品有其选择性，在细胞裂解物中除去干扰物质（如脱氧核糖核酸和酚类），蛋白酶被抑制后，蛋白质必须经过变性和溶解。最常见的裂解和溶解的溶液是根据 O' Farrell 的方法[7]，该裂解液含 2% NP-40，9mol/L 尿素，1% DTT 以及 0.8% 的载体两性电解质。

促溶剂（chaotrope），如尿素，是通过改变溶剂的参数发挥作用，在大多数 2D PAGE 过程中采用，然而，尿素引起的蛋白质变性增加了蛋白质之间的相互疏水作用，这个问题可以通过在溶液中添加硫脲素缓解。硫脲素是一种很强的变性剂，但不能单独使用，因为它的溶解度较低，在尿素-硫脲素裂解液中，必须添加一种去污剂以破坏脂类相互作用，非离子去污剂，如 3-［(3-胆胺丙基）二甲基氨］-1-丙烷磺酸酯（3-［(3-cholamidopropyl) dimethylammonio］-1-propane sulfonate，CHAPS）经常用到，另一种称为 sulfobetain 的去污剂也能产生很好的效果[8]。阴离子去污剂 SDS，在打断非共价蛋白质相互作用中有超强的功能，但它会干扰等电聚焦，SDS 可用于在加入裂解液稀释之前预溶解样品，裂解液用非离子去污剂把 SDS 从蛋白质中置换出来，这样蛋白质就处于溶解状态[9]。在还原二硫键的过程中，DTT 比巯基乙醇更优越，因为后者在碱性区会离子化，而破坏 pH 梯度。

双向电泳中最值得关注的问题是高分子质量蛋白、高度疏水性蛋白和碱性蛋白进入胶中受到限制[10, 11]。然而，利用新溶解剂、新 IPG（immobilized pH gradient，固定化的 pH 梯度）干胶条和新等电聚焦装置，对解决这些问题都很有帮助，如果对样品进行预分级分离效果更好[12~14]。

蛋白质分离

蛋白质混合物上样到有 pH 梯度的聚丙烯酰胺胶条上，给这个胶条高电压后，蛋白质会在电荷为零的 pH 处聚焦，利用平板胶的载体两性电解质（ampholyte）[15]或直接利用现成的 IPG 胶条[16]建立聚焦 pH 梯度。IPG 干胶条的最大优越性是提高实验重复性，使实验之间的结果具有可比性[17]，而采用 IPG，阴极的渗透飘移也不太显著，这样碱性蛋白的分辨率更好，聚焦时间可以延长，而使更大量的蛋白质样品上载到胶条上[18~20]。IPG 胶条的上样，既可用上样杯，也可通过胶条的再水化，再水化法在低电压（10～50V）可显著提高高分子质量蛋白质入胶[21]。

在第二向中，等电聚焦蛋白质根据分子质量相对大小在 SDS-PAGE 中分离，在这之前，IPG 胶条要在两种 SDS 溶液中分步骤平衡，第一步，在 DTT 中还原二硫键，第二步，利用碘乙酸铵（iodacetamide）碱化，以阻止再氧化。

质谱

在蛋白质组研究中，质谱已成为鉴定蛋白质的方法。在质谱中，分子的质量转化成气相离子，测量精度可达 50ppm[22]，最常用的两个离子化技术：一是 MALDI（matrix-assisted laser desorption ionization）[23]，另一是电子喷雾离子化[24]。

在 MALDI-TOF（time of flight，飞行时间）质谱法中[25]，蛋白质首先从胶上切割下来，用胰蛋白酶进行消化，得到的肽片段通过反相柱或微柱纯化，然后与介质（通常是 alpha-cyano-4-hydroxy cinnamic acid）一起固定在金属靶上，激光束打到靶上，介质分子吸收能量后分解蒸发，结果肽以离子化的形式进入气相[26]。MALDI 通常与 TOF 设备联用，测量粒子的质量，离子化的肽通过电场加速，根据它们到达探测器所用飞行时间（TOF）的不同，计算其质荷比。

大多数多肽的 50ppm 质量精度相当于 0.1 Da（同位素分辨率），收集的多肽质量，即多肽质量指纹，在理论上可用来检索胰蛋白酶酶解的蛋白质数据库[27]。根据质谱中匹配多肽的百分比、峰强度和比较观察、计算蛋白质的分子质量和 PI 来鉴别阳性蛋白质。整个过程如图 1 所示。

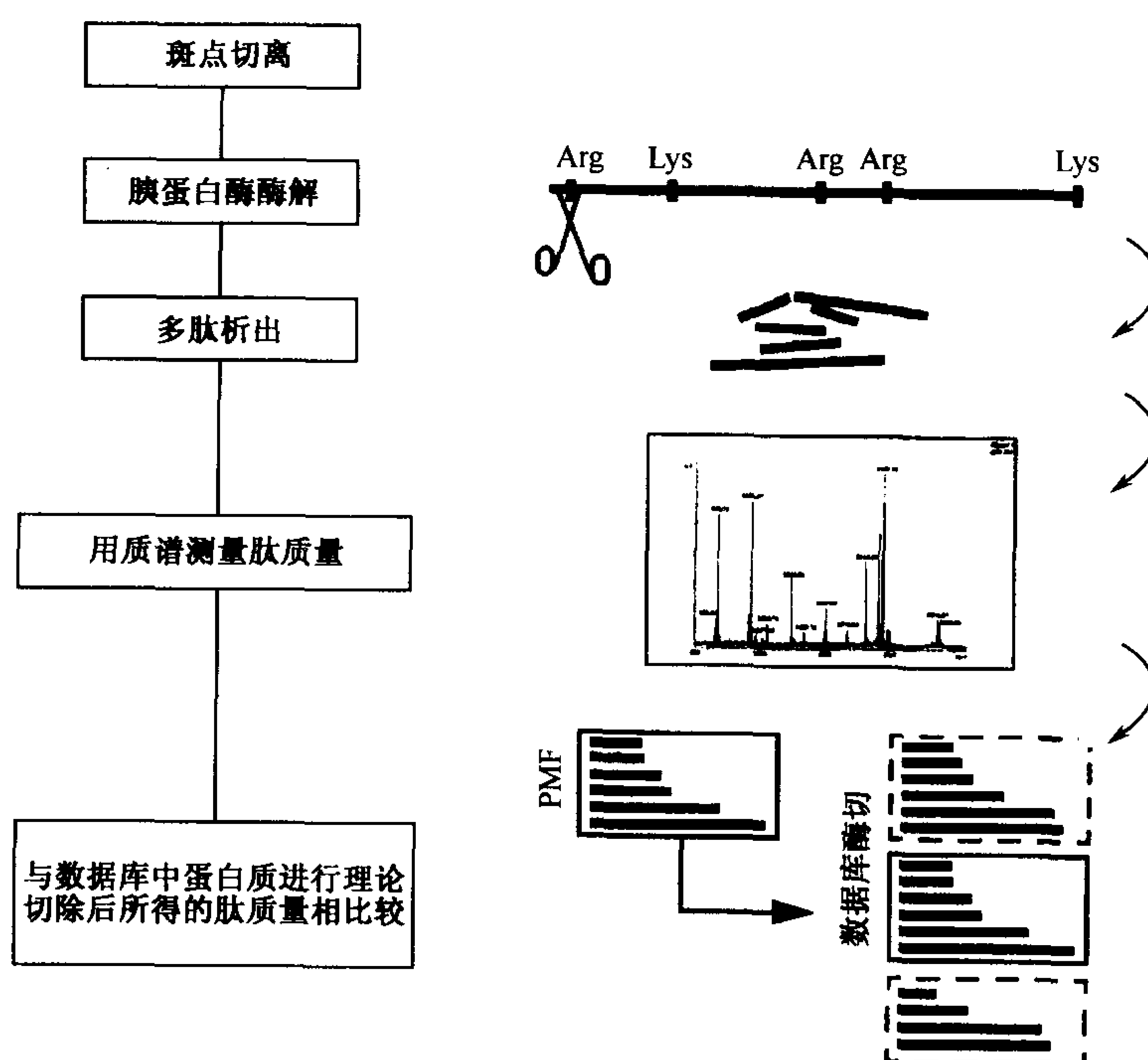


图 1　用质谱（MS）鉴别蛋白质。切离的斑点用胰蛋白酶酶解；多肽析出，除去盐分；样品用 MS 进行分析，得到多肽质量指纹（PMF）。通过在理论 PMF 数据库中查找所得的 PMF 来鉴别蛋白质。

在操作 MALDI-TOF MS 时，可以通过 PSD（postsource decay，后源衰退）[25]分析得到多肽序列标签（PST）[27]。后源衰退利用多肽离子化后要经过亚稳衰退的现象，说明同一速度的多肽片段有不同的质量，因而有不同的动能，不同的动能通过磁场反映出来，高能量片段比低能量片段穿透更深的磁场，因而会迟滞。一条多肽片段得到的谱带能推出多肽的氨基酸序列，因为断裂主要发生在肽键处。PST 能与蛋白质数据库匹配，由此可鉴别蛋白质来源[28]，大约 10 个氨基酸的 PST 可以提供唯一的鉴别，这些 PST 可从一种蛋白质的几个多肽片段中获得。

通过三重四极的 MS/MS 也可以获得 PST[29]，在四极中，离子纵向通过四根金属

棒以增加势能[30]，通过改变电压可以选择出一定质荷比（m/z）的离子，在第一个四极中，特定多肽被筛出；在第二个四极（含惰性气体）中，多肽由于碰撞导致断裂而片段化；在第三个四极中，记录碰撞导致断裂的谱带。

在MS/MS中，通常通过将多肽喷射入质谱仪中的电喷雾离子化完成多肽离子化[31]，这种离子化通过带电小滴的蒸发作用而达到，可用离子化器测量质量，它是通过保持离子在两电极间正弦运动的一定质荷比筛选离子[32]。

其他分离方法

先进的MS现在成为定量鉴别蛋白质的一种标准，它将现有细胞中所有蛋白质的混合物直接加载在质谱仪上，通过毛细管填充柱洗脱[4]。Shen和Smith发表过毛细管电泳与MS联用的综述[33]，这种方法可以通过同位素亲和标记物和MS/MS来定量分析[34]，不同样品中的蛋白质被非同位素标签或同位素标签（^{13}C或^{15}N）标记，质量不同的多肽在质谱中解析，峰密度用于计算两样品中蛋白的相对量。由于一种蛋白可产生很多不同多肽，所以，测量值在统计学上更正确。在另一方法中，用^{13}C或^{15}N作为单一N源，让细菌在不同环境中生长[35]，Tyers和Mann综述了用于研究蛋白质相互作用和蛋白质活性模型的定量MS方法[36]。

应用蛋白质组学

在蛋白质组学研究中，细菌是人们感兴趣的一种群体，因为它们的基因组相对较小，很多都已得到了序列，此外，可以收集不同时间、不同生长环境和不同微生物组成的样品材料。以下关于细菌蛋白质组学的论述，能够了解用蛋白质组分析方法可以获得相关信息。

微生物蛋白质组学

全基因组序列提供了在细胞水平通过实验鉴定表达蛋白的基础，但是，目前只有少数鉴定所有表达和潜在的修饰蛋白质研究。大肠杆菌蛋白质组用窄IPG胶条可以解析4950个白斑点，占理论蛋白质组的70%，但是，只有313个斑点能被MS鉴定出来，对应222种不同蛋白质，不能肯定是否有特异性蛋白质种类被遗漏[37]。

在确定全蛋白质组的尝试中，Ueberle等分析了无壁细菌肺炎支原体（*Mycoplasma pneumoniae*），比较了预测的688个可读框（ORF）和鉴定的基因产物[38]，蛋白质的鉴定效果很好，从450个蛋白质斑点中确定了224个基因，但是文章中明确指出2-D分离技术的局限性，因为只得到很少碱性pI蛋白质、分子质量大于100Da的蛋白质和含多跨膜片段的膜蛋白，他们同时分析了微生物菌体的细胞组分：细胞质蛋白、肝素结合蛋白、热稳定蛋白和核糖体蛋白[38]。等电点聚焦不适合分离核糖体蛋白，因为它们一般是碱性pI，即使用pH9.0～12.0的非线性pH值梯度，也只能鉴定52种预测核糖体蛋白中的12种，比较而言，通过单向SDS-PAGE后，用MS可以鉴定48种预测的核糖体蛋白。

支原体是已知最小独立生存的个体，肺炎支原体（*Mycoplasma pneumoniae*）基因组只有 570kb[39]，从基因组分析中可以得出 480 个预测的 ORF，这些蛋白质组涵盖了从 427 个蛋白质斑点中的 112 种蛋白质[40]。肺炎支原体蛋白质组之所以引起人们的注意，是因为人们期望从中能列出活细胞中必需的最小一套蛋白质，除了 17 种假设蛋白质外，发现了不同功能（细胞囊、新陈代谢、生物合成、转运、复制、转录和翻译）的蛋白质。这项研究的两个重要结果，是鉴定蛋白质的 Codon Adaptation Index 高于平均值，而碱性蛋白质和高分子质量蛋白质很少。将对数生长期细菌蛋白质的容量与稳定期的容量进行了比较，总之，蛋白质合成减少 42%，核糖体蛋白质相对丰度减少了 8 倍。目前，用 2-D PAGE 分离技术还不能完全覆盖微生物蛋白质组，但是，能得到其他研究不能预测的重要信息。

微生物组分蛋白质组学

通过蛋白质组学，可以在微生物菌体不同组分中鉴定蛋白质，如细胞质（细菌在机械破碎以及高速离心去除膜和核糖体后的可溶性蛋白）、核糖体、细胞器官，如菌毛、鞭毛、Ⅲ型分泌器官以及能被肌氨酰萃取纯化的革兰氏阴性细菌外膜复合体，此外，可以从培养物上清中得到分泌蛋白并对它们进行分析。

Butter 等[41]运用 2-D PAGE 详尽分析枯草芽孢杆菌（*Bacillus subtilis*）胞质体蛋白并用 MS 鉴定蛋白质，他们用高压破碎仪破碎细菌，然后离心除去细胞碎片，鉴定出 346 种蛋白质，几乎都是细胞质蛋白，这种方法对枯草芽孢杆菌非常有效。

用几乎相反的方法鉴定铜绿假单胞菌（*Pseudomonas aeruginosa*）的膜亚蛋白质组[42]，只检测到一小部分主要胞质蛋白 GroEL，表明此方法可以去掉胞质蛋白。作者观察到用两性离子洗涤剂 CHAPS 和氨基硫酸三甲铵乙内酯来溶解蛋白质存在的差异，后者可以溶解更多的蛋白质。

革兰氏阴性细菌的外膜能用去污剂肌氨酰（sarkosyl）提取得到外膜复合体（OMC）[43]，详细研究了衣原体外膜复合体 OMC，衣原体（*Chlamydia*）是重要人类病原，它们是专性细胞内革兰氏阴性细菌，有两阶段发育周期，在胞外是有传染性的原体（EB），直径约 0.3μm，在胞内是无传染性能复制的网状体（RB），直径约 1μm，EB 附着寄主细胞，诱导自身被吸收进特化的液泡，称为衣原体内含物，EB 进入细胞后重组成 RB，后者通过二分分裂而增殖，当接近胞内末期，RB 转换成 EB，最终，有传染性的新一代 EB 释放出来，寄主细胞被裂解[44]。

在单向凝胶中，沙眼衣原体（*Chlamydia trachomatis*）的 OMC 显示三个主要带（Omp2、Omp3 和 MOMP），而肺炎衣原体（*Chlamydia pneumoniae*）有额外的约 100kDa 带，这些含多种蛋白带的现象称为外膜蛋白多样性[45]，在肺炎衣原体基因组中，发现了 21 个多样性外膜蛋白（Pmp）的基因[46]。

Grimwood 等[47]以每个 Pmps 抗体，通过免疫杂交显示大部分 *pmp* 基因得到表达。Vandahl 等[48]用 2-DPAGE 分离肺炎衣原体的外膜复合体蛋白，发现在研究的 pH 范围内所有 7 种 Pmp（总共有 10 种外膜蛋白在整个微生物体的蛋白质组中被鉴定）都存在于 OMC 中（图 2），这说明表达的 Pmp 位于外膜复合体[49]。

Pmp 有自身转运蛋白质的结构特点[50]，C 端部分在膜外形成一个 β 折叠桶，N 端的被转运乘客结构域为平行 β 折叠结构[51]。有趣的是，蛋白质组显示 3 个蛋白质（Pmp6，Pmp20 和 Pmp21）被降解，通过 N 端序列鉴定发现，切割位点位于自身转运结构域和乘客结构域之间[48]。

一些致病的革兰氏阴性细菌有分泌器官，能在细菌胞质里合成的蛋白质注入到真核寄主细胞里（Ⅲ型分泌），基因编码的Ⅲ型分泌器官在物种间有相似性，却没有共同的特性去定义分泌效应蛋白，衣原体有Ⅲ型分泌基因[52]，Vandahl 等[52]鉴定了肺炎衣原体（*C. pneumoniae*）EB 的 YscC、YscN、YscL 和 LcrE，YscN 和 YscL 位于细胞质膜上，YscC 定位在外膜上，衣原体接触寄主细胞细胞质或小泡膜时，LcrE（即 CopN）控制效应蛋白的释放[53]。与这一模式类似，在 OMC 中只有 Ysc 存在（图 2），说明衣原体Ⅲ型分泌装置像其他细菌一样地装配[49]，LcrE 并不是一种外膜整合蛋白，在沙眼衣原体中该蛋白由Ⅲ型装置分泌，定位于包含体膜上[54]。

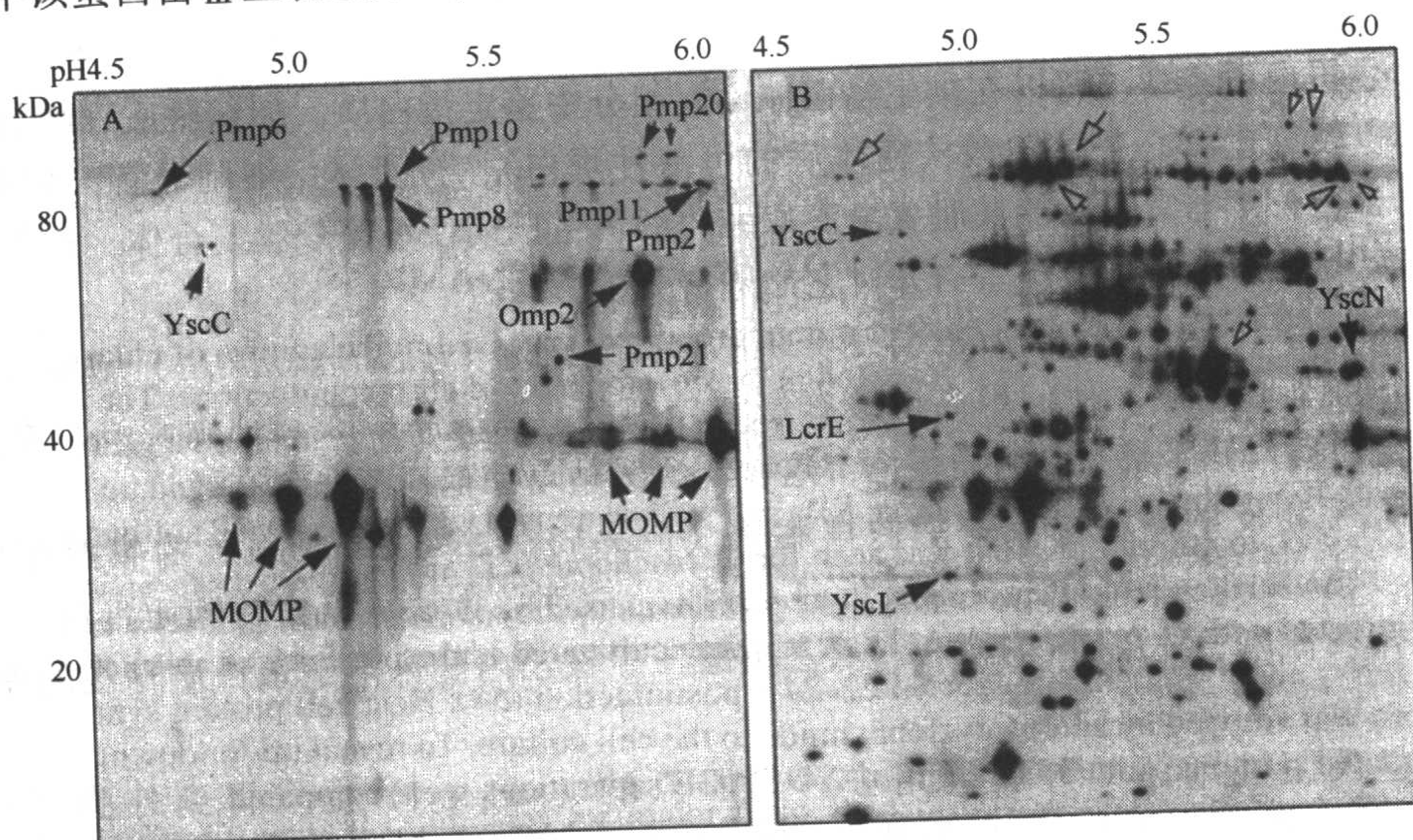


图 2　银染色的 IPG4-7，2-D PAGE 凝胶负载了肺炎衣原体（*C. pneumoniae*）的 OMC。(A) 1.5mg 纯化的 EB　B. 250μg 肺炎衣原体的 EB。标记蛋白用从 EB 凝胶中胰蛋白酶酶解的蛋白质斑点通过 MS 进行鉴定。(B) 空心箭头标出阳性的 Pmp 蛋白。

用蛋白质组学研究蛋白动力学

γ 干扰素在控制衣原体感染方面是一种强有力的免疫调节剂，并对慢性感染的发展也有抑制作用。沙眼衣原体的 A、B、C 血清型对 γ 干扰素十分敏感，对被感染 A 型沙眼衣原体的 HeLa 细胞，用 γ 干扰素处理后的荧光免疫显微法观察到非典型的大 RB，并与抗 MOMP 抗体形成轻微染色[55~57]，这与 D 型和 L2 型沙眼衣原体中不同。

为了显示经 γ 干扰素处理后在蛋白质组上的不同，将感染了沙眼衣原体 A、D 和

L2 的 HeLa 细胞在感染 22～24 小时后，培养在有或无 γ 干扰素和^{35}S 标记的蛋氨酸中[58]，寄主细胞的蛋白质合成由在细胞培养物中加入放线菌酮而终止，采用双向 PAGE 放射自显影比较法可以揭示蛋白质的表达变化。

γ 干扰素负调节沙眼衣原体蛋白

对沙眼衣原体 A 的双向 PAGE 分析表明，经 γ 干扰素处理后，包括 MOMP 在内的一些蛋白被负调节，而 GroEL 的表达水平不变，这些被负调节的蛋白，包括 C1pC 蛋白酶和Ⅰ型二磷酸果糖醛缩酶（Fba）[56,57]。这个观察结果与早先的结论相吻合[55,59,60]，但不仅仅限于重要的免疫源，这种依赖 γ 干扰素的负调节在沙眼衣原体 D 和 L2 中没观察到[56,57]。

γ 干扰素诱导沙眼衣原体色氨酸合成酶

沙眼衣原体 D 基因组中的色氨酸操纵子（trp operon），包括色氨酸合成酶 A（trp A)和色氨酸合成酶 B（trp B）两个亚单位，还有色氨酸阻抑物（trp R）[61]。双向 PAGE 显示沙眼衣原体 A、D 和 L2 对 γ 干扰素的反应是由于 Trp A 和 Trp B 的强烈诱导[57]。与 Trp B 不同的是，沙眼衣原体 A 和 Trp A 的分子质量大大低于沙眼衣原体 D 和 L2 的 Trp A 分子质量[56,57]，这是因为 1bp 的缺失导致移码阅读和早熟终止密码子切除了约 70 个氨基酸序列（～7.7kDa）。沙眼衣原体 C 中有一个与沙眼衣原体 A 中相同的缺失。沙眼衣原体 B 在基因组中有一段包括色氨酸操纵子在内的序列缺失[62]，血清型 A-C 中 Trp AB 的切断或缺失，可能使这些血清型对 γ 干扰素介导的色氨酸降解更敏感，使这些血清型对寄主细胞感染更加持久并导致长期感染[57]。

分泌蛋白

在整个细胞内的生命周期中，衣原体处于包含体内，包含体外有包膜，这是一种修饰过的吞噬体膜。衣原体与寄主细胞的联系可能由分泌蛋白介导，在包含体的胞内发育阶段可以修饰寄主细胞，而且，分泌蛋白是 MHC Ⅰ 型抗原的天然候选者，因此，可作为开发疫苗的靶标。

以蛋白质组学为基础的一种方法，用于比较^{35}S 标记衣原体感染细胞全部溶解物（WLIC）中的衣原体蛋白、提纯 RB 和 EB 被^{35}S 标记的衣原体蛋白。在 WLIC 中有分泌性蛋白存在，而在纯 RB 和 EB 中并不存在（图 3）。候选分泌蛋白通过质谱分析法（MS）检测出来，可以认为，衣原体蛋白酶或类蛋白酶体（proteasomelike）的活性因子（CPAF）被分泌到寄主细胞中[63]，并在细胞质中降解寄主细胞的转录因子 RFX5 和 USF-1[65]，RFX5 和 USF-1 是Ⅰ型和Ⅱ型 MHC 抗原存在的必要条件。

Zhong 等[63]的结果被 Shaw 等[66]的研究证实和深化，沙眼衣原体和肺炎衣原体 CPAF 的 N 端和 C 端通过 2-D 胶在 WLIC 中检测到，但 RB 和 EB 中却没有，这表明：沙眼衣原体 A、D 和 L2 及肺炎衣原体分泌 CPAF 是在生活的中期和末期。CPAF 的序列在沙眼衣原体、肺炎衣原体和鹦鹉热衣原体（*Chlamydia psittaci*）中很保守，也就是说可能所有衣原体细菌都能在其生活周期中分泌 CPAF。在结合双向 PAGE 和脉冲追

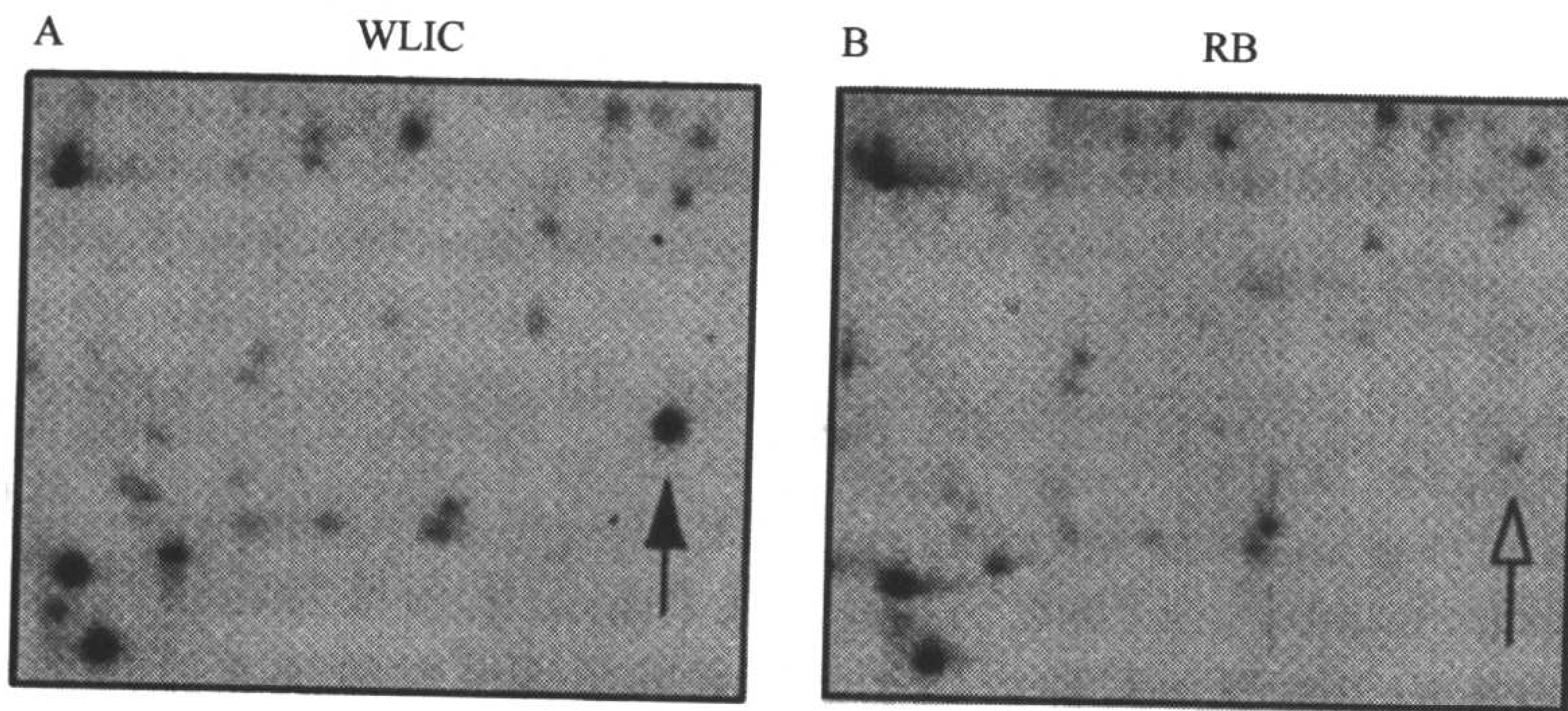


图 3　从感染肺炎衣原体的细胞全裂解液（WLIC）和从肺炎衣原体的 RB 中，用 2D PAGE 分离蛋白的放射自显影图。寄主细胞的蛋白质合成由放线菌酮进行终止。感染的细胞由 ^{35}S-蛋氨酸标记 36～38 小时；RB 在 38 小时时被纯化。箭头所指为 CPAF 的 N 端（CPN1016 N-term）。

踪的研究中，CPAF 的两个片段在寄主中表现得相当稳定，表明这两个片段所进化出的氨基酸序列在寄主细胞中不易降解。衣原体分泌抑制Ⅰ型和Ⅱ型 MHC 抗原出现的抗体，并且使自身也不致成为 MHCⅠ型抗原表达的靶位点，由于蛋白质组学的建立和发展，快速筛选分泌蛋白有助于鉴定衣原体分泌的其他蛋白。

2-D 胶电泳数据库

研究 DNA 序列（欧洲生物信息研究所/国家生物技术信息中心）和蛋白质数据（SwissProt）有一个秩序井然的数据库系统，并有镜像连接多个网址。2-D 胶数据库必须包括凝胶图和蛋白质相关数据，Swiss 2-D PAGE 是 1993 年建立的双向电泳数据库。它包括了从酿酒酵母（*Saccharomyces cerevisiae*）、大肠杆菌和盘基网柄菌（*Dictyostelium discoideum*）在内，直到人类和鼠类组织的 24 种参考图谱。

通过使用 ExPASy 的 Make 2-D-DB 软件，使得蛋白质点阵的视图与区分更加容易[67]，当点击 2-D 图谱中某图上的一点时，蛋白质的信息（等电点、分子质量、登录号、连接到 SwissProt 数据库）和含此蛋白质的 2D 图谱都会列出来。通过蛋白质登录号的搜索，也可以在 2-D 图谱数据库中找到该蛋白质。而且，在 SwissProt 中还有链接与蛋白质 2D 图谱库相连。Make 2-D DB 软件，在目前被包括 ExPASy Swiss 2-D PAGE 在内的 7 个场所应用，如果想查阅所有 2-D PAGE 数据库和服务内容，请访问 http://www.expasy.org。

新版 Make 2-D DB 在功能上有很大升级，它是一个相关性数据库，在其中可以设计很多高级查询项。由于 2-D 胶图像和蛋白斑点之间的复杂相关性，对于 2-D 图像数据的持续添加，系统被分散，每个研究组有独立运行各自的数据库。使用 2-D DB 时，一个查询指令能够被传达到所有 2-D PAGE 数据库，采用 Make 2-D DB 使得 ExPASy 能够管理各研究小组的数据库，从而确保数据的长期使用。

Max Planck Institute for Infection Biology 研究所维护一个包含 *Mycobacterium*、

Mycoplasma、*Chlamydia*、*Francisella*、*Helicobacter* 和 *Borrelia* 在内的蛋白质组信息的互联网数据库[68]，该研究所已经开发出一个数据输入软件 Topspot，并运行了一个包括自己和其他小组的 2-D PAGE 数据在内的中央数据库[68]，这是有很大潜在价值的数据库。从两株毒性菌株 *Mycobacterium tuberculesis*（H37RV and Erdman）到两株非毒性菌株 *Mycobacterium bovis BCG*（Chicago and Copenhagen）的蛋白质组都已比较过，而且总数超过 800 个点（spot）都已被鉴定[69]，发现毒性菌株间有 16 种蛋白质的差异，毒性菌株 H37Rv 与非毒性菌株有 25 种蛋白质的差异[70]，毒株与非毒株的蛋白差异可能与毒性有关，希望这些可能的致病蛋白能为肺结核的药物治疗提供靶标。

在 *Helicobacter pylori* 中，已经鉴别出 126 种不同蛋白质[71]，这些蛋白质包括绝大多数具有代表性的蛋白质类型，包括转录和翻译因子。*Helicobacter pyloris* 能在极其酸性条件下生长，这对它能够在胃中感染寄主很重要，在酸性环境生长的过程中，检测到上调节表达的毒力因子蛋白，早先描述并不认为与耐酸性有关的蛋白 Htr A 等以及早先已知的“酸性”蛋白（Vac A）被检测到。强抗原是疫苗的候选物[70]，通过对病人血清的免疫印迹（immunoblotting）分析，检测到新潜在毒力因子，早先描述过的和新抗原都有所发现。值得一提的是，所有早先描述为在对动物的免疫研究中起保护作用的 *H. pylori* 蛋白均是高丰度蛋白[71]。

将加里螺旋体菌（*Borrelia garinii*）蛋白组与一种称为 Lyme 疏螺旋体病患者血清进行免疫印迹，从加里螺旋体菌蛋白组中筛选抗原的实验[72]，监测了不同病态的 20 种血清中与 217 种蛋白质的反应，发现了 65 种抗原，其中有 20 种已被鉴别，并且有 3-磷酸甘油醛脱氢酶和 ATP-binding-cassette transporter 寡肽透性酶两种是新发现。

两种数据库系统都可以将鉴别方法和 MS 数据整合在一起，截至 2003 年 4 月，互联网上已有 44 个 2D PAGE 的数据库（http://www.expasy.org/ch2d/2d-index.html），其中大多数与疾病和微生物有关。

蛋白质组学作为基因组学的补充

虽然还不可能在全蛋白质组中鉴别出一套完整的蛋白质，但已经清楚，蛋白质组学可以通过很多途径补充与完善基因组学，例如，用 Edman 降解法鉴定切下的蛋白斑点[73]，对未被注释的蛋白质进行鉴定[38,74]、寻找蛋白片段和鉴别分泌蛋白[48,66]都是蛋白质组学补充与完善基因组学的例子。2-D PAGE 的主要优势体现在对蛋白质转换（turnover）上，在这里对细菌的脉冲标记和脉冲追踪能鉴别出特定蛋白质的合成与加工[56,66]。对在不同生长条件下合成蛋白质的脉冲标记和 2-D PAGE 成像的比较，是一种鉴定诱导蛋白和下调蛋白的有用技术，对蛋白质组学的彻底了解是有价值的。但在研究方法达到快速、高效、能完全研究蛋白质组之前，与 MS 联用的 2-D PAGE 系统，仍然是研究蛋白质表达与加工的最佳选择。与 MS 联用的毛细管电泳相比，MS 与 2-D PAGE 联用的方法为人们提供了一个可以通过肉眼识别、判断图像的选择，它的作用不可低估。

（刘子铎，吴天福 译）

参考文献

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995; 269:496–512.
2. National Center for Biotechnology Information. Prominent Organisms Taxonomy/List. November 13, 2003. http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/org.html.
3. Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000; 405: 827–836.
4. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. Evaluation of two- dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci USA 2000; 97:9390–9395.
5. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A sampling of the yeast proteome. Mol Cell Biol 1999; 19:7357–7368.
6. Wilkins MR, Pasquali C, Appel RD, et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology (NY) 1996; 14:61–65.
7. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. J Biol Chem 1975; 250:4007–4021.
8. Rabilloud T, Adessi C, Giraudel A, Lunardi J. Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 1997; 18:307–316.
9. Dunn MJ, Bradd SJ. Separation and analysis of membrane proteins by SDS-polyacrylamide gel electrophoresis. Methods Mol Biol 1993; 19:203–210.
10. Santoni V, Molloy M, Rabilloud T. Membrane proteins and proteomics: un amour impossible? Electrophoresis 2000; 21:1054–1070.
11. Adessi C, Miege C, Albrieux C, Rabilloud T. Two-dimensional electrophoresis of membrane proteins: a current challenge for immobilized pH gradients. Electrophoresis 1997; 18:127–135.
12. Herbert B. Advances in protein solubilisation for two-dimensional electrophoresis. Electrophoresis 1999; 20:660–663.
13. Rabilloud T. Use of thiourea to increase the solubility of membrane proteins in two-dimensional electrophoresis. Electrophoresis 1998; 19:758–760.
14. Gorg A, Obermaier C, Boguth G, et al. The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 2000; 21:1037–1053.
15. Righetti PG, Gianazza E. New developments in isoelectric focusing. J Chromatogr 1980; 184: 415–456.
16. Bjellqvist B, Ek K, Righetti PG, et al. Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. J Biochem Biophys Methods 1982; 6:317–339.
17. Corbett JM, Dunn MJ, Posch A, Gorg A. Positional reproducibility of protein spots in two-dimensional polyacrylamide gel electrophoresis using immobilised pH gradient isoelectric focusing in the first dimension: an interlaboratory comparison. Electrophoresis 1994; 15:1205–1211.
18. Sanchez JC, Rouge V, Pisteur M, et al. Improved and simplified in-gel sample application using reswelling of dry immobilized pH gradients. Electrophoresis 1997; 18:324–327.
19. Gorg A, Obermaier C, Boguth G, Weiss W. Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures. Electrophoresis 1999; 20:712–717.
20. Gorg A, Obermaier C, Boguth G, et al. The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 2000; 21:1037–1053.

21. Zuo X, Speicher DW. Quantitative evaluation of protein recoveries in two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 2000; 21:3035–3047.
22. Jensen ON, Podtelejnikov AV, Mann M. Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. Anal Chem 1997; 69: 4741–4750.
23. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 1988; 60:2299–2301.
24. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989; 246:64–71.
25. Gevaert K, Vandekerckhove J. Protein identification methods in proteomics. Electrophoresis 2000; 21:1145–1154.
26. Zenobi R, Knochenmuss R. Ion formation in MALDI mass spectrometry. Mass Spectrom Rev 1998; 17:337–366.
27. Mann M, Hojrup P, Roepstorff P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom 1993; 22:338–345.
28. Wilkins MR, Gasteiger E, Sanchez JC, Appel RD, Hochstrasser DF. Protein identification with sequence tags. Curr Biol 1996; 6:1543–1544.
29. Andersen JS, Mann M. Functional genomics by mass spectrometry. FEBS Lett 2000; 480:25–31.
30. Yost RA, Boyd RK. Tandem mass spectrometry: quadrupole and hybrid instruments. Methods Enzymol 1990; 193:154–200.
31. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989; 246:64–71.
32. Cooks RG, Glish GL, Kaiser RE, McLuckey SA. Ion trap mass spectrometry. Chem Eng News 1991; 69:26–41.
33. Shen Y, Smith RD. Proteomics based on high-efficiency capillary separations. Electrophoresis 2002; 23:3106–3124.
34. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 1999; 17:994–999.
35. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci USA 1999; 96:6591–6596.
36. Tyers M, Mann M. From genomics to proteomics. Nature 2003; 422:193–197.
37. Tonella L, Hoogland C, Binz PA, Appel RD, Hochstrasser DF, Sanchez JC. New perspectives in the *Escherichia coli* proteome investigation. Proteomics 2001; 1:409–423.
38. Ueberle B, Frank R, Herrmann R. The proteome of the bacterium *Mycoplasma pneumoniae*: comparing predicted open reading frames to identified gene products. Proteomics 2002; 2: 754–764.
39. Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995; 270:397–403.
40. Wasinger VC, Pollack JD, Humphery-Smith I. The proteome of *Mycoplasma genitalium*. Chaps-soluble component. Eur J Biochem 2000; 267:1571–1582.
41. Buttne K, Bernhardt J, Scharf C, et al. A comprehensive two-dimensional map of cytosolic proteins of *Bacillus subtilis*. Electrophoresis 2001; 22:2908–2935.
42. Nouwens AS, Cordwell SJ, Larsen MR, et al. Complementing genomics with proteomics: the membrane subproteome of *Pseudomonas aeruginosa* PAO1. Electrophoresis 2000; 21:3797–3809.
43. Caldwell HD, Kromhout J, Schachter J. Purification and partial characterization of the major outer membrane protein of *Chlamydia trachomatis*. Infect Immun 1981; 31:1161–1176.
44. Birkelund S. The molecular biology and diagnostics of *Chlamydia trachomatis*. Dan Med Bull 1992; 39:304–320.

45. Knudsen K, Madsen AS, Mygind P, Christiansen G, Birkelund S. Identification of two novel genes encoding 97- to 99-kilodalton outer membrane proteins of *Chlamydia pneumoniae*. Infect Immun 1999; 67:375–383.
46. Kalman S, Mitchell W, Marathe R, et al. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet 1999; 21:385–389.
47. Grimwood J, Olinger L, Stephens RS. Expression of *Chlamydia pneumoniae* polymorphic membrane protein family genes. Infect Immun 2001; 69:2383–2389.
48. Vandahl BB, Pedersen AS, Gevaert K, et al. The expression, processing and localization of polymorphic membrane proteins in *Chlamydia pneumoniae* strain CWL029. BMC Microbiol 2002; 2:36.
49. Vandahl BB, Christiansen G, Birkelund S. 2D-page analysis of the *Chlamydia pneumoniae* outer membrane complex. In: Schachter J, Christiansen G, Clarke IN, et al. (eds). Proceedings of the 10th Symposium on Human Chlamydial Infections. June 16–21, International Chlamydia Symposium, San Francisco, CA, 2002, pp. 547–550.
50. Henderson IR, Lam AC. Polymorphic proteins of *Chlamydia* spp—autotransporters beyond the Proteobacteria. Trends Microbiol 2001; 9:573–578.
51. Birkelund S, Christiansen G, Vandahl BB, Pedersen AS. Are the Pmp proteins parallel β-helices? In: Schachter J, Christiansen G, Clarke IN, et al. (eds). Proceedings of the 10th Symposium on Human Chlamydial Infections. June 16–21, International Chlamydia Symposium, San Francisco, CA, 2002, pp. 551–554.
52. Vandahl BB, Birkelund S, Demol H, et al. Proteome analysis of the *Chlamydia pneumoniae* elementary body. Electrophoresis 2001; 22:1204–1223.
53. Rockey DD, Lenart J, Stephens RS. Genome sequencing and our understanding of chlamydiae. Infect Immun 2000; 68:5473–5479.
54. Fields KA, Hackstadt T. Evidence for the secretion of *Chlamydia trachomatis* CopN by a type III secretion mechanism. Mol Microbiol 2000; 38:1048–1060.
55. Beatty WL, Byrne GI, Morrison RP. Morphologic and antigenic characterization of interferon gamma-mediated persistent *Chlamydia trachomatis* infetion in vitro. Proc Natl Acad Sci USA 1993; 90:3998–4002.
56. Shaw AC, Christiansen G, Birkelund S. Effects of interferon gamma on *Chlamydia trachomatis* serovar A and L2 protein expression investigated by two-dimensional gel electrophoresis. Electrophoresis 1999; 20:775–780.
57. Shaw AC, Christiansen G, Roepstorff P, Birkelund S. Genetic differences in the *Chlamydia trachomatis* tryptophan synthase a-subunit can explain variations in serovar pathogenesis. Microbes Infect 2000; 2:581–592.
58. Shaw AC, Gevaert K, Demol H, et al. Comparative analysis of *Chlamydia trachomatis* serovar A, D and L2. Proteomics 2002; 2:164–186.
59. Beatty WL, Belanger TA, Desai AA, Morrison RP, Byrne GI. Tryptophan depletion as mechanism for gamma-interferin-mediated chlamydial persistence. Infect Immun 1994; 62:3705–3711.
60. Beatty WL, Morrison RP, Byrne GI. Reactivation of persistent *Chlamydia trachomatis* infection in cell culture. Infect Immun 1995; 63:199–205.
61. Stephens RS, Kalman S, Lammel C, et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 1998; 282:754–759.
62. Stephens RS (Ed.) Chlamydia. Intracellular Biology, Pathogenesis, and Immunity, Washington, DC: ASM Press, 1999, pp. 9–27.
63. Zhong G, Fan P, Ji H, Dong F, Huang Y. Identification of a chlamydial protease-like activity factor responsible for the degradation of host transcription factors. J Exp Med 2001; 193:935–942.

64. Fan P, Dong F, Huang Y, Zhong G. *Chlamydia pneumoniae* secretion of a protease-like activity factor for degrading host cell transcription factors is required for major histocompatibility complex antigen expression. Infect Immun 2002; 70:345–349.
65. Zhong G, Fan T, Liu L. *Chlamydia* inhibits interferon-γ inducible major histocombability complex class II expression by degradation of upstream stimulatory factor 1. J Exp Med 1999; 189: 1931–1938.
66. Shaw AC, Vandahl BB, Larsen MR, et al. Characterization of a secreted *Chlamydia* protease. Cell Microbiol 2002; 4:411–424.
67. Hoogland C, Baujard V, Sanchez JC, Hochstrasser DF, Appel RD. Make2ddb: a simple package to set up a two-dimensional electrophoresis database for the World Wide Web. Electrophoresis 1997; 18:2755–2758.
68. Eifert T, Büttner S. Proteome 2D-PAGE database, November 26, 2003. http://www.mpiib-berlin.mpg.de/2D-PAGE/.
69. Jungblut PR, Schaible UE, Mollenkopf HJ, et al. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. Mol Microbiol 1999; 33:1103–1117.
70. Cash P. Proteomics in medical microbiology. Electrophoresis 2000; 21:1187–1201.
71. Jungblut PR, Bumann D, Haas G, et al. Comparative proteome analysis of *Helicobacter pylori*. Mol Microbiol 2000; 36:710–725.
72. Jungblut PR, Grabher G, Stoffler G. Comprehensive detection of immunorelevant *Borrelia garinii* antigens by two-dimensional electrophoresis. Electrophoresis 1999; 20:3611–3622.
73. Wasinger VC, Humphery-Smith I. Small genes/gene-products in *Escherichia coli* K-12. FEMS Microbiol Lett 1998; 169:375–382.
74. Shaw AC, Larsen MR, Roepstorff P, Christiansen G, Birkelund S. Identification and characterization of a novel *Chlamydia trachomatis* reticulate body protein. FEMS Microbiol Lett 2002; 212:193–202.

索　引